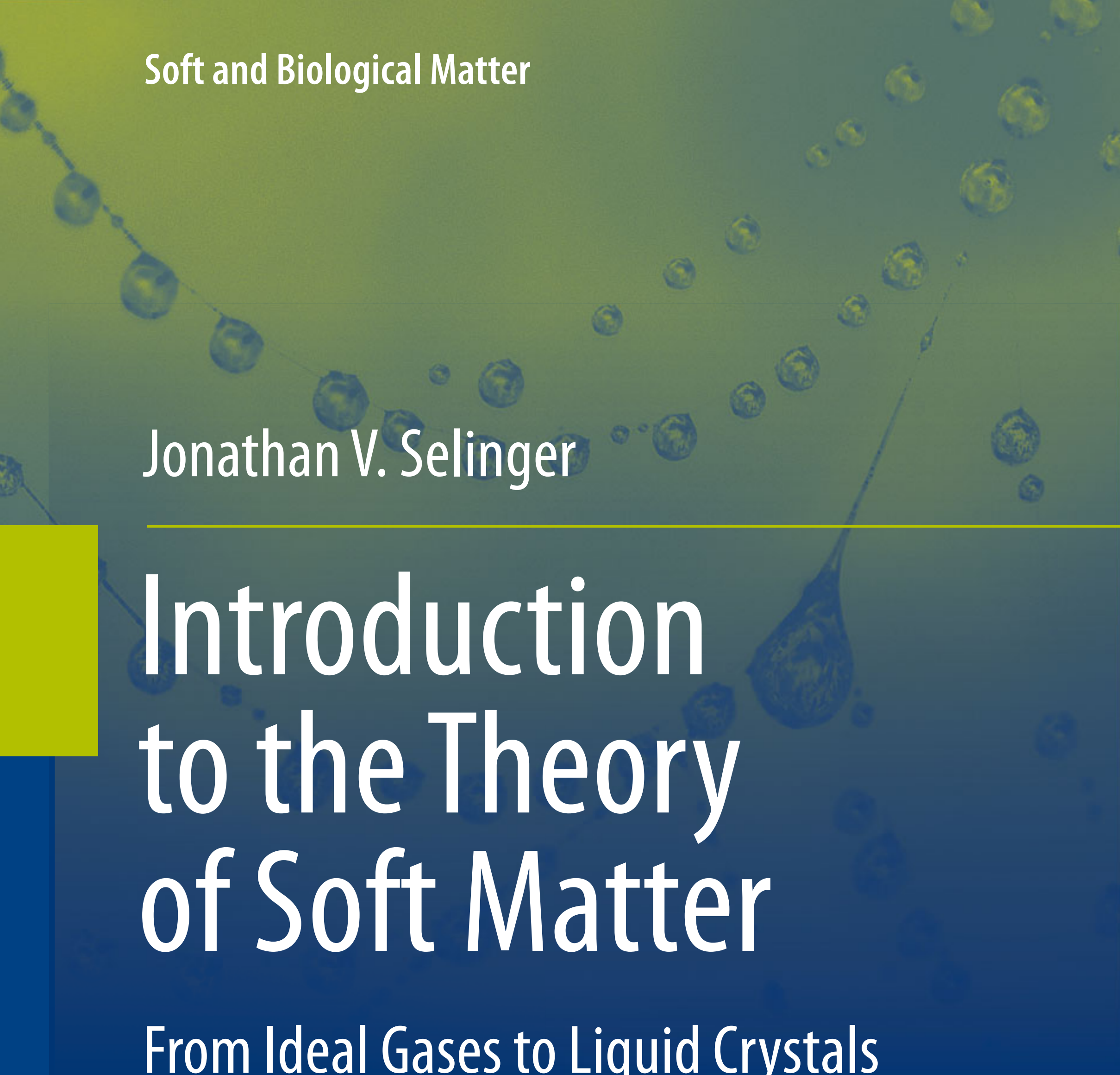


Soft and Biological Matter

Jonathan V. Selinger

Introduction to the Theory of Soft Matter

From Ideal Gases to Liquid Crystals

Springer

Soft and Biological Matter

"Soft and Biological Matter" is a series of authoritative books covering established and emergent areas in the realm of soft matter science, including biological systems spanning from the molecular to the mesoscale. It aims to serve a broad interdisciplinary community of students and researchers in physics, chemistry, biophysics and materials science.

Pure research monographs in the series as well as those of more pedagogical nature, will emphasize topics in fundamental physics, synthesis and design, characterization and new prospective applications of soft and biological matter systems. The series will encompass experimental, theoretical and computational approaches.

Both authored and edited volumes will be considered.

More information about this series at http://www.springer.com/series/10783

Jonathan V. Selinger

Introduction to the Theory of Soft Matter

From Ideal Gases to Liquid Crystals

Springer

Jonathan V. Selinger
Chemical Physics Interdisciplinary Program
Liquid Crystal Institute
Kent State University
Kent, OH
USA

For some of the figures in this book, there is an interactive electronic version in Wolfram Computable Document Format (CDF). These interactive figures are available at this URL: http://www.springer.com/cda/content/document/cda_downloaddocument/Selinger+Interactive+Figures.zip?SGWID=0-0-45-1509169-p177545420.

ISSN 2213-1736 ISSN 2213-1744 (electronic)
Soft and Biological Matter
ISBN 978-3-319-21053-7 ISBN 978-3-319-21054-4 (eBook)
DOI 10.1007/978-3-319-21054-4

Library of Congress Control Number: 2015946076

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media
(www.springer.com)

To Robin, for the collaboration of my life

Preface

This book is based on a course on “Soft Matter,” which I teach for first-year graduate students in the Chemical Physics Interdisciplinary Program at the Liquid Crystal Institute of Kent State University. Students come into this program from several undergraduate majors—including physics, chemistry, materials science, chemical, or electrical engineering—and from many countries. The purpose of my course is to teach this diverse group of students about the statistical physics aspects of liquid crystals and other soft materials. At the same time, other professors teach the students about other aspects of these materials—including chemistry, optics, and design of devices. The students can then combine all of these scientific disciplines in their Ph.D. dissertation research.

In teaching this course, I have found that there are several excellent undergraduate-level books that describe the experimental phenomena of soft materials. There are also many excellent graduate-level books on the theoretical physics of these materials. However, I believe that students need a guide to help them make the transition between these levels—a basic introduction to theoretical physics, explaining the concepts of symmetry, broken symmetry, and order parameters; phases and phase transitions; mean-field theory; and the mathematics of variational calculus and tensors.

I have written this book to meet that need. In particular, I have tried to make it useful for two types of students: First, there are students going into theoretical research. This book will help them to progress toward studying more advanced topics and reading more advanced books on theoretical physics. Second, there are students going into other disciplines, such as experimental physics, chemistry, and engineering. These students may never plan to study more advanced theory, but this book will prepare them to understand theoretical seminars, read theoretical articles, and collaborate with theoretical colleagues.

To serve this diverse audience of students, I have taken two steps. First, I have intentionally written the book in an informal, conversational style. I find that this style is accessible to students from a wide range of backgrounds—from different scientific fields as well as different nationalities.

Second, as a technological innovation, the book is accompanied by a set of "interactive figures" (available at http://www.springer.com/cda/content/document/cda_downloaddocument/Selinger+Interactive+Figures.zip?SGWID=0-0-45-1509169-p177545420). Some of these figures allow readers to change parameters and see what happens to a graph, some allow readers to rotate a plot or other graphics in 3D, and some do both. The interactive figures help students to develop their intuition for the physical meaning of equations. I strongly urge all readers to download and try them while reading the book. They are in a Wolfram Computable Document Format (CDF) file, which can be opened with Mathematica or with Wolfram's free CDF player (available at https://www.wolfram.com/cdf-player). The player has versions for Windows, Mac, and Linux (although unfortunately not for iOS or Android).

I would like to thank all the students in my classes, as well as my research students, for their feedback as I developed the concepts for this book. I particularly thank Thanh-Son Nguyen for his careful reading of the first draft.

I gratefully acknowledge National Science Foundation Grants DMR-1106014 and 1409658, which supported my research while I wrote this book. Furthermore, I thank the Liquid Crystal Institute and Kent State University for inviting me to join their faculty 10 years ago, and for their consistent support of my research and teaching.

Kent, OH, USA Jonathan V. Selinger
April 2015

Contents

Chapter 1
Toy Model

Abstract This chapter presents a simple "toy model" to illustrate the concepts of energy, entropy, and free energy. In this model, multiple microstates are grouped together into a single macrostate through a process of coarse-graining. The system tends to go into the macrostate that minimizes the free energy. At low temperature, this is the ordered state with the lowest energy. At high temperature, it is the disordered state with the highest entropy or multiplicity.

To see a world in a grain of sand,
And a heaven in a wild flower,
Hold infinity in the palm of your hand,
And eternity in an hour.
...
William Blake, Auguries of Innocence

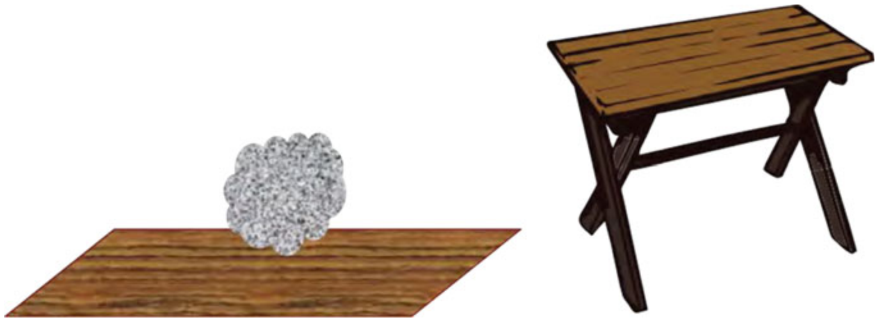
Physicists are always looking for the Big Ideas—ideas that are fundamentally simple in concept but can be applied to understand many phenomena throughout nature. In my field of statistical mechanics and phase transitions, the Big Idea is the balance between order and disorder, a balance that is controlled by temperature. In this chapter, I would like to introduce you to this Big Idea through a *toy model,* i.e., a model which is not a serious theory but is just meant as a simple way to illustrate a basic point. I want to see the world not in a grain of sand, but in a speck of dust.[1]

Let us begin with the situation shown in Fig. 1.1. Here, a speck of dust can have two states: it can be either on the floor or on the table. Each of these states has some energy E_{floor} or E_{table}. Based on the Boltzmann distribution,[2] these two states have the probabilities $p_{\text{floor}} = (e^{-E_{\text{floor}}/k_B T})/Z$ and $p_{\text{table}} = (e^{-E_{\text{table}}/k_B T})/Z$, where k_B

[1] Disclaimer: Of course, this is not a real model of the physics of dust. It is just a toy!.

[2] I am assuming that you have already seen the Boltzmann distribution in a previous course. If you have not, you should consult a textbook on thermal physics or statistical thermodynamics, as listed at the end of the chapter.

J.V. Selinger, *Introduction to the Theory of Soft Matter*,
Soft and Biological Matter, DOI 10.1007/978-3-319-21054-4_1



	Floor	Table
Energy	E_{floor}	E_{table}
Probability	$(e^{-E_{\text{floor}}/k_B T})/Z$	$(e^{-E_{\text{table}}/k_B T})/Z$

Fig. 1.1 Toy model for one table, with the corresponding energies and probabilities

is Boltzmann's constant,[3] T is the temperature, and $Z = e^{-E_{\text{floor}}/k_B T} + e^{-E_{\text{table}}/k_B T}$ is the *partition function* that makes the probabilities add up to 1.

Which state is more likely? Let us compare the two probabilities: The speck of dust is more likely to be on the floor if

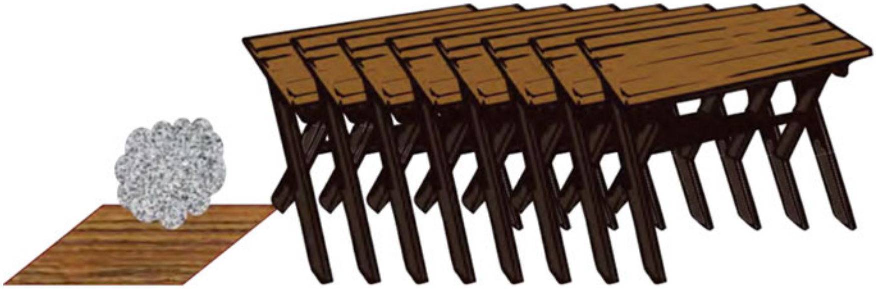
$$\begin{aligned} p_{\text{floor}} &> p_{\text{table}}, \\ \frac{e^{-E_{\text{floor}}/k_B T}}{Z} &> \frac{e^{-E_{\text{table}}/k_B T}}{Z}, \\ -\frac{E_{\text{floor}}}{k_B T} &> -\frac{E_{\text{table}}}{k_B T}, \\ E_{\text{floor}} &< E_{\text{table}}. \end{aligned} \tag{1.1}$$

Hence, the speck is more likely to be on the floor than the table if the floor energy is lower than the table energy. That is not a big surprise!

Now let us consider a slightly more interesting problem. Suppose there are N tables, where N is some large number, as shown in Fig. 1.2. The probability of being on the floor is $(e^{-E_{\text{floor}}/k_B T})/Z$. The probability of being on a *particular* table is $(e^{-E_{\text{table}}/k_B T})/Z$. The probability of being on *any* of the N tables is $N(e^{-E_{\text{table}}/k_B T})/Z$. Here, the partition function that normalizes the probabilities must be $Z = e^{-E_{\text{floor}}/k_B T} + N e^{-E_{\text{table}}/k_B T}$.

In this new problem, let us compare the probabilities again. The speck is more likely to be on the floor than on *any* of the tables if

[3] Why do I use the subscript B in Boltzmann's constant? Well, my PhD advisor wrote it that way, and now I cannot help myself.



	Floor	Any table
Energy	E_{floor}	E_{table}
Probability	$(e^{-E_{\text{floor}}/k_BT})/Z$	$N(e^{-E_{\text{table}}/k_BT})/Z$
Entropy	0	$k_B \log N$
Free energy	E_{floor}	$E_{\text{table}} - k_BT \log N$

Fig. 1.2 Toy model for many tables, with the corresponding energies, probabilities, entropies, and free energies

$$\begin{aligned} p_{\text{floor}} &> p_{\text{any table}}, \\ \frac{e^{-E_{\text{floor}}/k_BT}}{Z} &> \frac{Ne^{-E_{\text{table}}/k_BT}}{Z}, \\ -\frac{E_{\text{floor}}}{k_BT} &> -\frac{E_{\text{table}}}{k_BT} + \log N, \\ E_{\text{floor}} &< E_{\text{table}} - k_BT \log N. \end{aligned} \tag{1.2}$$

This little calculation shows that we should not just compare the *energy* E_{floor} and E_{table}. Instead, we should compare some adjusted energy that takes into account the multiplicity of the states. The adjustment (without the factor of T) is called the *entropy*, and the adjusted energy is called the *free energy*. For the group of all table states, the entropy is

$$S_{\text{table}} = k_B \log N, \tag{1.3}$$

and the free energy is

$$F_{\text{table}} = E_{\text{table}} - TS_{\text{table}} = E_{\text{table}} - k_BT \log N. \tag{1.4}$$

Likewise, because there is only one floor state, the entropy of the floor is

$$S_{\text{floor}} = k_B \log(1) = 0, \tag{1.5}$$

and the free energy of the floor

$$F_{\text{floor}} = E_{\text{floor}} - TS_{\text{floor}} = E_{\text{floor}} - k_B T \log(1) = E_{\text{floor}}. \tag{1.6}$$

Hence, the speck is more likely to be on the floor than on any of the tables if

$$F_{\text{floor}} < F_{\text{table}}. \tag{1.7}$$

From this toy model, I think we learn three general lessons. First, we see the importance of grouping states together. If we care about whether the speck is on *any* table, and we do not care *which* table, then it is appropriate to group all of the table states together. In this grouping, we are treating a collection of microstates as a single macrostate. This macrostate has an energy E and a multiplicity N (and hence an entropy $S = k_B \log N$). This grouping is our first example of the concept of *coarse-graining,* which will be fundamental throughout statistical mechanics.

Second, we see the concept of free energy, which combines the energy and entropy of a macrostate into the quantity $F = E - TS$. This quantity is essential for understanding what macrostate is most likely to occur at a nonzero temperature T.

Third, we see that the relative importance of energy and entropy depends on temperature. At low temperature, energy is the dominant part of the free energy, and the system is most likely to go into whatever macrostate has the lowest energy. That is usually a single, special state, which might be called an *ordered* state. By contrast, at high temperature, entropy is the dominant part of the free energy, and the system is most likely to go into whatever macrostate has the highest entropy. That is usually a big collection of many microstates. They each have a high energy, so they are individually unlikely, but there are *a lot* of them! That collection might be called a *disordered* state.

Now, dear students, I imagine that some of you might have an objection. You might say "What do you mean by defining the free energy *of a state*? I have already learned the definition of free energy in a class on thermal physics or statistical thermodynamics, and it was different! In that class, we did not talk about the free energy *of a state*; we just talked about *THE* free energy. It was defined in terms of the partition function Z as"

$$F = -k_B T \log Z. \tag{1.8}$$

"So what's going on here?!?"

I am glad you asked that question. Let us look at Eq. (1.8) for *THE* free energy, and rewrite it as

$$e^{-F/k_B T} = Z = e^{-E_{\text{floor}}/k_B T} + e^{-E_{\text{table }1}/k_B T} + \cdots + e^{-E_{\text{table }N}/k_B T}. \tag{1.9}$$

Now we group the N table microstates together into a single macrostate, and rewrite the equation as

$$e^{-F/k_BT} = Z = e^{-E_{\text{floor}}/k_BT} + Ne^{-E_{\text{table}}/k_BT} \tag{1.10}$$

$$= e^{-F_{\text{floor}}/k_BT} + e^{-F_{\text{table}}/k_BT}. \tag{1.11}$$

Look at this: When we combine many microstates together in the partition function, we get the free energy of the combined macrostate. We can keep combining macrostates together to make super-macrostates, and combining super-macrostates to make super-duper-macrostates, and each of them has a free energy. When we eventually finish combining all possible states together, we reach *THE* free energy of the system. That is coarse-graining!

At this point, I think we have learned everything we can from this toy model, and we need to move on to something more physical. Onward!

Further Reading

This book assumes that you are already familiar with the basic principles of thermal physics or statistical thermodynamics, particularly the Boltzmann distribution. If you need to learn these principles, some textbooks are

1. R. Baierlein, *Thermal Physics* (Cambridge, 1999)
2. C. Kittel, R. Kroemer, *Thermal Physics* (Freeman, 1980)
3. R. Reif, *Fundamentals of Statistical and Thermal Physics* (Waveland, 2008)
4. D.V. Schroeder, *An Introduction to Thermal Physics* (Addison-Wesley, 1999)

A more advanced book, with an unusual point of view, is

5. J.P. Sethna, *Statistical Mechanics: Entropy, Order Parameters, and Complexity* (Oxford, 2006)

It is available from the the author's website http://sethna.lassp.cornell.edu/statistical_mechanics_entropy_order_parameters_and_complexity.

Chapter 2
Ising Model for Ferromagnetism

Abstract This chapter presents the Ising model for ferromagnetism, which is a standard simple model of a phase transition. Using the approximation of mean-field theory, the free energy is minimized, and hence the magnetization is calculated, as a function of temperature and applied field. This calculation demonstrates some fundamental concepts in statistical mechanics, including spontaneous symmetry breaking, an order parameter with magnitude and direction, first- and second-order phase transitions, and a critical point characterized by critical exponents. In the rest of the book, these concepts will be applied to the theory of soft matter.

In this chapter, I will introduce the Ising model for ferromagnetism. This is probably the single most commonly studied model in statistical mechanics—one might say that the Ising model is to statistical mechanics as the fruit fly is to genetics. As you will see, the Ising model shows the essential concept of how the balance between energy and entropy leads to a phase transition.

Historical note: This model of ferromagnetism was developed in 1924 by Professor Wilhelm Lenz and his graduate student Ernst Ising. As far as I know, it is the only case in the history of science where the work was named after the student, not after the professor. I promise you that will *NEVER* happen again!

2.1 Model

In the Ising model, we consider a lattice of magnetic moments, as shown in Fig. 2.1. (In this figure, I have drawn a two-dimensional (2D) square lattice, but in general, we could have any lattice structure in any dimension.) On each lattice site, the local magnetic moment is represented by a "spin," drawn as an arrow in the figure. We assume that the spin has just two possible states, either pointing up or pointing down. Mathematically, we represent the spin at site i by the variable $\sigma_i = \pm 1$. In this notation, $+1$ means that the spin is pointing up, and -1 means that it is pointing down.

The energy for the Ising model includes two contributions: the interaction between neighboring spins and the effect of an applied magnetic field on each individual spin. The interaction between neighboring spins tends to induce parallel alignment of the

J.V. Selinger, *Introduction to the Theory of Soft Matter*,
Soft and Biological Matter, DOI 10.1007/978-3-319-21054-4_2

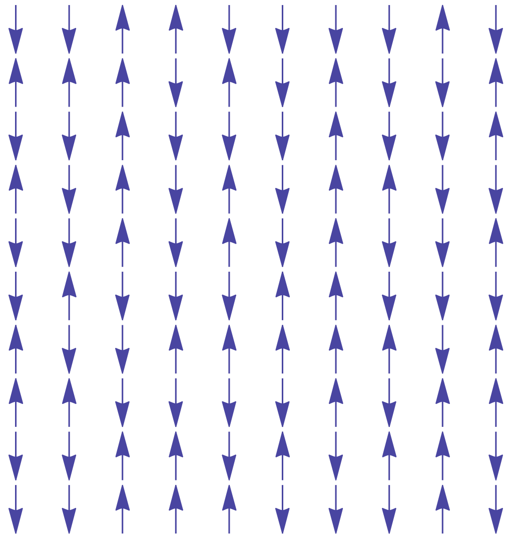


Fig. 2.1 Example of the Ising model on a 2D square lattice. Each *arrow* is a "spin," which represents a magnetic moment that can point either *up* or *down*

neighbors, so it should be favorable (negative) when the neighbors are both $+1$ or both -1, and unfavorable (positive) when the neighbors are $+1$ next to -1. Hence, for each pair of neighbors i and j, the interaction energy can be written as $-J\sigma_i\sigma_j$, where J is a *positive* coefficient giving the interaction strength.[1] If the magnetic field h is pointing up, it favors each spin pointing up; if the field is pointing down, it favors each spin pointing down.[2] Hence, for each site i, the field energy can be written as $-h\sigma_i$. Putting these pieces together, the total energy for the system becomes

$$E = -J\sum_{\langle i,j\rangle}\sigma_i\sigma_j - h\sum_i \sigma_i. \tag{2.1}$$

Note the indices on these two sums. In the first sum, the angle brackets $\langle i, j\rangle$ represent *nearest-neighbor* pairs (for example, north-south or east-west on the 2D square lattice); the sum is taken over all nearest-neighbor pairs. By comparison, the second sum is taken over all individual sites i, which are each affected by the magnetic field.

Do not worry about the edges of the system; we will neglect them in our discussion. That is a reasonable approximation if the system is very large, so that only a tiny fraction of the sites are on the surface.

Our goal is now to calculate how much magnetic order is in the system. Suppose there are N spins in the lattice, with $N_\uparrow$ spins pointing up and $N_\downarrow$ spins pointing down,

[1] The parameter J is sometimes called the "exchange constant," for reasons based on the quantum mechanics of magnetism.

[2] To be precise, h is proportional to the magnetic field H, scaled by the magnetic moment μ per spin; people often disregard the factor of μ and refer to h as a field.

so that $N = N_{\uparrow} + N_{\downarrow}$. The total magnetic moment of the system is $\mu(N_{\uparrow} - N_{\downarrow})$, where μ is the magnetic moment of each spin, and the largest possible magnetic moment is μN. Hence, it is natural to define the “magnetic order parameter” or “magnetization” M as the expectation value of the magnetic moment relative to the largest possible magnetic moment,

$$M = \left\langle \frac{N_{\uparrow} - N_{\downarrow}}{N} \right\rangle. \tag{2.2}$$

Hence, we want to calculate M as a function of the interaction strength J, the magnetic field h, and the temperature T.

Note that M can assume values from -1 to 1. The absolute value of M indicates the magnitude of magnetic order. If $|M|$ is close to 0, then the system is highly disordered, with approximately half of the spins pointing up and half pointing down. By comparison, if $|M|$ is close to 1, the system is highly ordered, with almost all of the spins pointing in the same direction. The positive or negative signs of M indicate the direction of magnetic order—if it is positive, then the net order is pointing up; if it is negative, then the net order is pointing down. In further chapters, we will see that these are very general features of order parameters; they always show the magnitude and direction of order.

2.2 Non-interacting Spins

As a first step, just for practice, let us do the calculation for non-interacting spins with $J = 0$. In this case, we can solve for M exactly, and it will help us get ready for the *much* harder problem of interacting spins with $J > 0$.

In this section, I will present two ways to solve the problem of non-interacting spins. The first approach is a standard solution, which you have probably seen in courses on thermal physics or statistical thermodynamics. The second approach is a more interesting solution in terms of energy and entropy.

2.2.1 Standard Solution

For the standard solution, we begin with the partition function

$$Z = \sum_{\text{states}} e^{-E_{\text{state}}/k_B T}. \tag{2.3}$$

Here, a “state” refers to the full list of the values of the spins σ_1, σ_2, etc., and the energy of a state is $E = -h \sum_i \sigma_i$ in the non-interacting model. Hence, the partition function becomes

$$Z = \sum_{\sigma_1=\pm 1} \sum_{\sigma_2=\pm 1} \cdots \sum_{\sigma_N=\pm 1} e^{(h/k_B T)\sum_i \sigma_i}. \tag{2.4}$$

Because the energy consists of separate terms for each spin, with no interactions, the sum factorizes into

$$Z = \left[\sum_{\sigma_1=\pm 1} e^{(h/k_B T)\sigma_1}\right]\left[\sum_{\sigma_2=\pm 1} e^{(h/k_B T)\sigma_2}\right]\cdots\left[\sum_{\sigma_N=\pm 1} e^{(h/k_B T)\sigma_N}\right]. \tag{2.5}$$

Because each of those factors is identical, the partition function for N spins factorizes into the product of single-spin partition functions,

$$Z = Z_1^N, \tag{2.6}$$

where

$$Z_1 = \sum_{\sigma_1=\pm 1} e^{(h/k_B T)\sigma_1} = e^{+h/k_B T} + e^{-h/k_B T}. \tag{2.7}$$

For each site, the probabilities of pointing up or down are

$$p_\uparrow = \frac{e^{+h/k_B T}}{e^{+h/k_B T} + e^{-h/k_B T}}, \tag{2.8}$$

$$p_\downarrow = \frac{e^{-h/k_B T}}{e^{+h/k_B T} + e^{-h/k_B T}}. \tag{2.9}$$

Hence, the expectation value of any single spin is

$$\langle \sigma_i \rangle = (+1)p_\uparrow + (-1)p_\downarrow = \frac{e^{+h/k_B T} - e^{-h/k_B T}}{e^{+h/k_B T} + e^{-h/k_B T}} = \tanh\left(\frac{h}{k_B T}\right). \tag{2.10}$$

Likewise, because all the spins are identical, the magnetic order parameter is

$$M = \tanh\left(\frac{h}{k_B T}\right). \tag{2.11}$$

Figure 2.2 shows a plot of this result for M as a function of $h/k_B T$. Note that M is zero at $h = 0$, i.e., this non-interacting model has no magnetic order without a field. Moreover, M saturates at its maximum value of $+1$ when $h \to +\infty$, and at -1 when $h \to -\infty$. We might ask: How large of a magnetic field is required to induce an order parameter of, say, 75 % of its maximum value? From Eq. (2.11), we see that it occurs at $h/k_B T = \tanh^{-1}(0.75) \approx 1$, or in other words, at $h \approx k_B T$. Hence, the temperature determines how sharply M saturates as a function of h. Only a small field is required at low temperature, but a much larger field is required at high temperature. A related question is: How strongly does M respond to a small applied

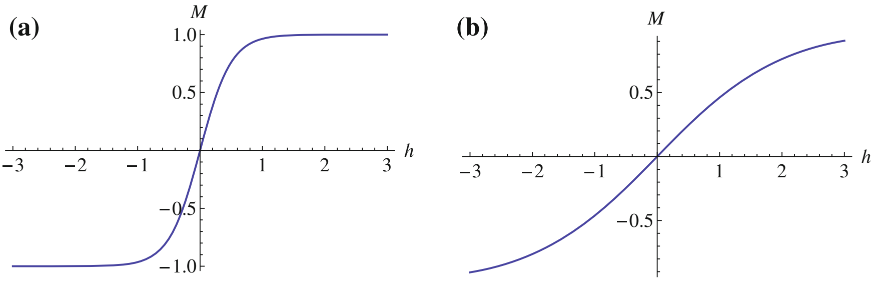


Fig. 2.2 Magnetic order parameter of the *non-interacting* Ising model, as a function of h. **a** For $k_B T = 0.5$. **b** For $k_B T = 2$ (Interactive version at http://www.springer.com/cda/content/document/cda_downloaddocument/Selinger+Interactive+Figures.zip?SGWID=0-0-45-1509169-p177545420.)

field, i.e., what is the derivative of M with respect to h at $h = 0$? This derivative is called the *susceptibility* χ, and it can be calculated as

$$\chi \equiv \left.\frac{\partial M}{\partial h}\right|_{h=0} = \frac{1}{k_B T}. \tag{2.12}$$

Hence, M responds extremely sensitively to small h in the limit of low temperature.

2.2.2 Solution in Terms of Energy and Entropy

I would now like to present a different solution to the same problem of non-interacting spins in terms of energy and entropy. It will, of course, give the same answer!

As we recall from the toy model of dust on a table, a key concept in statistical mechanics is classifying microstates into macrostates. In the Ising model, each microstate refers to a particular configuration of up and down spins. Let us now classify these microstates according to the value of M. For example, if we have a system of 100 spins, we can have one macrostate with $M = 0$ (50 up, 50 down), another macrostate with $M = 0.02$ (51 up, 49 down), etc. In general, if we have a system of N spins, with $N_\uparrow$ up and $N_\downarrow$ down, then

$$\begin{aligned} N_\uparrow + N_\downarrow &= N, \\ N_\uparrow - N_\downarrow &= NM. \end{aligned} \tag{2.13}$$

These equations imply

$$\begin{aligned} N_\uparrow &= N p_\uparrow = N\left(\frac{1+M}{2}\right), \\ N_\downarrow &= N p_\downarrow = N\left(\frac{1-M}{2}\right). \end{aligned} \tag{2.14}$$

We can now ask: How many microstates correspond to the macrostate with a certain value of M? In other words, in how many ways can we divide N total spins into $N_\uparrow$ spins pointing up and $N_\downarrow$ spins pointing down, where $N_\uparrow$ and $N_\downarrow$ are given by Eq. (2.14)? This is a standard combinatorial problem, which you might have studied in a class on probability and statistics. The answer is given by the binomial coefficient

$$\#\text{ of microstates} = \binom{N}{N_\uparrow} = \frac{N!}{N_\uparrow! N_\downarrow!} \tag{2.15}$$

Hence, the entropy associated with that value of M is

$$S(M) = k_B[\log(N!) - \log(N_\uparrow!) - \log(N_\downarrow!)]. \tag{2.16}$$

For large N, we can approximate $N!$ by Stirling's formula

$$\log(N!) \approx N \log N - N, \tag{2.17}$$

and likewise for $\log(N_\uparrow!)$ and $\log(N_\downarrow!)$. Hence, the entropy simplifies to

$$\begin{aligned} S(M) &= k_B[N \log N - N_\uparrow \log N_\uparrow - N_\downarrow \log N_\downarrow] \qquad (2.18)\\ &= -N k_B[p_\uparrow \log p_\uparrow + p_\downarrow \log p_\downarrow]\\ &= -N k_B\left[\left(\frac{1+M}{2}\right)\log\left(\frac{1+M}{2}\right) + \left(\frac{1-M}{2}\right)\log\left(\frac{1-M}{2}\right)\right]. \end{aligned}$$

What about the energy? For the non-interacting system, the energy is just the sum of the field energy terms for each of the N spins:

$$E(M) = (N_\uparrow)(-h) + (N_\downarrow)(+h) = -NhM. \tag{2.19}$$

Hence, the free energy as a function of M is

$$\begin{aligned} F(M) &= E(M) - TS(M) \qquad (2.20)\\ &= -NhM\\ &\quad + N k_B T\left[\left(\frac{1+M}{2}\right)\log\left(\frac{1+M}{2}\right) + \left(\frac{1-M}{2}\right)\log\left(\frac{1-M}{2}\right)\right]. \end{aligned}$$

Note that the free energy is proportional to the number of spins N, as it should be. This implies that the free energy is *extensive;* if we double the number of spins, then we double the free energy. Hence, we can factor out N to obtain the free energy per spin:

$$\frac{F(M)}{N} = -hM + k_B T\left[\left(\frac{1+M}{2}\right)\log\left(\frac{1+M}{2}\right) + \left(\frac{1-M}{2}\right)\log\left(\frac{1-M}{2}\right)\right]. \tag{2.21}$$

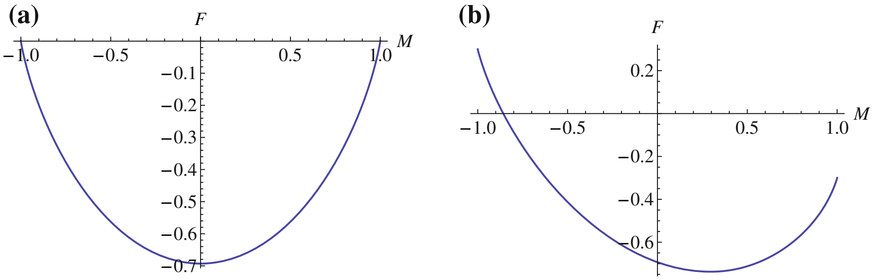


Fig. 2.3 Free energy of the *non-interacting* Ising model, as a function of M. **a** For $h/k_BT = 0$. **b** For $h/k_BT = 0.3$ (Interactive version at http://www.springer.com/cda/content/document/cda_downloaddocument/Selinger+Interactive+Figures.zip?SGWID=0-0-45-1509169-p177545420.)

What does this free energy look like, as a function of M? Well, the function involves two parameters, h and k_BT. It would be easier to analyze the shape of the function if it had only one parameter. For that reason, we divide both sides of the equation by k_BT to obtain

$$\frac{F(M)}{Nk_BT} = -\left(\frac{h}{k_BT}\right)M + \left(\frac{1+M}{2}\right)\log\left(\frac{1+M}{2}\right) + \left(\frac{1-M}{2}\right)\log\left(\frac{1-M}{2}\right). \tag{2.22}$$

Now we can plot it for several values of the single parameter h/k_BT. The results are shown in Fig. 2.3. Note that the minimum depends on the value of h/k_BT. If $h/k_BT = 0$, the minimum is exactly at $M = 0$. If $h/k_BT > 0$, the minimum shifts to positive values of M. As $h/k_BT \to \infty$, the minimum shifts toward $M = 1$. Likewise, if $h/k_BT < 0$, the minimum is at negative values of M, and it approaches $M = -1$ as $h/k_BT \to -\infty$.

To find the minimum algebraically, we just calculate the derivative and set it equal to zero:

$$\frac{\partial}{\partial M}\left(\frac{F(M)}{Nk_BT}\right) = -\frac{h}{k_BT} + \frac{1}{2}\log\left(\frac{1+M}{1-M}\right) = 0. \tag{2.23}$$

The solution is

$$M = \tanh\left(\frac{h}{k_BT}\right), \tag{2.24}$$

which is exactly the same as Eq. (2.11)! This solution is already plotted in Fig. 2.2, and we discussed it there. Hence, these solutions are consistent.

What do we learn from this second version of the solution? Well, we see that order parameter M is controlled by a competition between energy and entropy. The entropy favors the macrostate with $M = 0$, because this macrostate has the most microstates. By contrast, the energy favors the largest possible value of M aligned with the field h (positive M if $h > 0$, negative M if $h < 0$). By minimizing the free energy, we can find the equilibrium value of M. We will use that concept throughout this book.

2.3 Interacting Spins

Now let us consider the case where the spins are interacting. In other words, we are back to the Ising energy of Eq. (2.1) with the interaction strength $J > 0$. We would like to calculate the magnetic order parameter M as a function of J, h, and T in this case. Unfortunately, this problem is much harder than the non-interacting spins. It is not just harder in the sense that I need to look up a tricky integral, or that I have to get Mathematica to calculate something numerically. It is harder in the sense that it consists of a huge number of variables that are all coupled together. Instead of the non-interacting partition function of Eq. (2.4), we now have

$$Z = \sum_{\sigma_1=\pm 1} \sum_{\sigma_2=\pm 1} \cdots \sum_{\sigma_N=\pm 1} e^{(J/k_B T) \sum_{\langle i,j \rangle} \sigma_i \sigma_j + (h/k_B T) \sum_i \sigma_i}. \tag{2.25}$$

If we try to factorize this partition function into separate terms for each spin, as in Eq. (2.5), it does not work! The J term couples spin #1 to its neighbors, and those spins to their neighbors, and so forth, until *all* the spins are coupled together. For that reason, we cannot solve N copies of the same single-spin problem; we must solve a single N-spin problem. For a macroscopic system, N is a very large number, perhaps 6×10^{23}. It is almost always impossible to solve a problem like that exactly.

Although we cannot solve the problem exactly, there is a very useful approximation called *mean-field theory*, which provides a lot of insight into the behavior. In this approximation, we neglect the correlations between neighboring spins, and assume that they are each fluctuating independently with the same statistical distribution. In the following two sections, I will explain this approximation to you in two different ways.

2.3.1 Mean-Field Theory in Terms of Energy and Entropy

For a first approach to mean-field theory for the Ising model, let us continue to work in terms of energy and entropy. By analogy with the non-interacting model, we will classify the microstates into macrostates according to their order parameter M. We already worked out the entropy as a function of M in Eq. (2.18), and the field energy as a function of M in Eq. (2.19), and we will continue to use those results. Hence, we just need to work out an expression for the interaction energy as a function of M.

For any particular microstate with specific spins σ_1, σ_2, ... σ_N, the interaction energy is given by

$$E_{\text{int}} = -J \sum_{\langle i,j \rangle} \sigma_i \sigma_j. \tag{2.26}$$

Clearly it does not just depend on M; it depends on all the spins. Can we at least calculate the expectation value of that energy? The expectation value is given by

$$\langle E_{\text{int}} \rangle = -J \sum_{\langle i,j \rangle} \langle \sigma_i \sigma_j \rangle. \tag{2.27}$$

It involves correlations between neighboring spins σ_i and σ_j, which we do not know.

In mean-field theory, we now make a *HUGE* approximation. We *neglect* the correlations between neighboring spins, and write

$$\langle \sigma_i \sigma_j \rangle \approx \langle \sigma_i \rangle \langle \sigma_j \rangle. \tag{2.28}$$

If you have any experience with statistics, you will not like that line; you will complain to me that it is not true. You are right; it is not true; it is an approximation! We cannot yet tell whether it is a good approximation or a bad approximation. As you will see, it actually works surprisingly well, much better than we have any right to expect.

If you believe the mean-field approximation, then we can write

$$\langle E_{\text{int}} \rangle \approx -J \sum_{\langle i,j \rangle} \langle \sigma_i \rangle \langle \sigma_j \rangle. \tag{2.29}$$

Because all the spins are identical, they all have the same expectation value: $\langle \sigma_i \rangle = \langle \sigma_j \rangle = M$ for all sites i and j. Hence, each term in the sum reduces to M^2. How many terms are in the sum? Well, let us suppose that each site on the lattice has q nearest neighbors; q is called the *coordination number* of the lattice. In general, q depends on the lattice type and dimensionality. For the 2D square lattice shown in Fig. 2.1, we have $q = 4$. If there are N sites on the lattice, and each site interacts with q neighbors, you might expect that there are Nq pairs of nearest neighbors in the sum. Unfortunately, this argument double-counts the pairs, i. e. it counts σ_1 as a neighbor of σ_2, and σ_2 as a neighbor of σ_1. When we eliminate the double-counting, the number of nearest-neighbor pairs in the sum is $\frac{1}{2}Nq$. Hence, our mean-field approximation for the interaction energy is

$$\langle E_{\text{int}} \rangle \approx -\frac{1}{2} N J q M^2. \tag{2.30}$$

We now construct the free energy $F = \langle E \rangle - TS$ that includes the mean-field approximation for the interaction energy, along with the field energy and the entropy,

$$\begin{aligned} \frac{F(M)}{Nk_B T} = &- \left(\frac{Jq}{2k_B T}\right) M^2 - \left(\frac{h}{k_B T}\right) M \\ &+ \left(\frac{1+M}{2}\right) \log\left(\frac{1+M}{2}\right) + \left(\frac{1-M}{2}\right) \log\left(\frac{1-M}{2}\right). \end{aligned} \tag{2.31}$$

This free energy depends on two parameters, $Jq/k_B T$ and $h/k_B T$. To understand its behavior, we need to consider various cases.

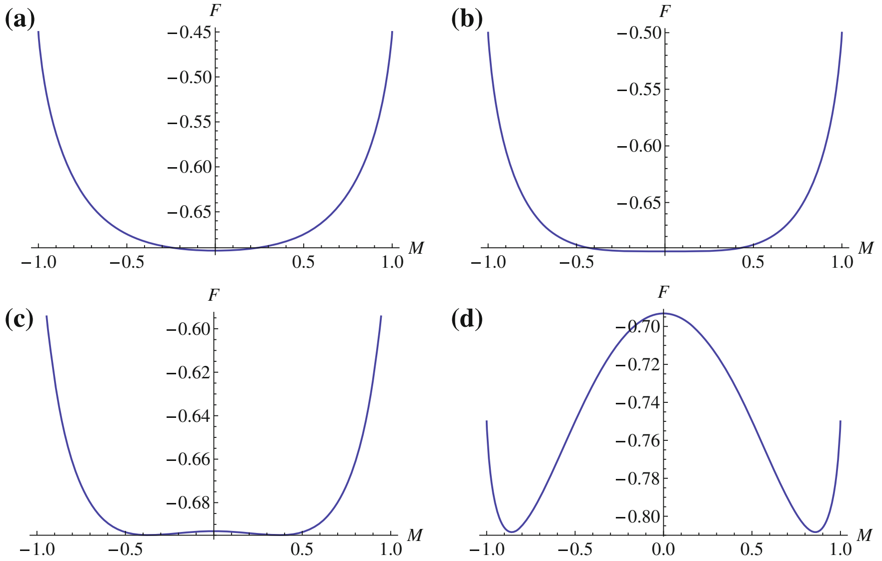


Fig. 2.4 Free energy of the *interacting* Ising model, with $h/k_BT = 0$, as a function of M. **a** For $Jq/k_BT = 0.9$. **b** For $Jq/k_BT = 1$. **c** For $Jq/k_BT = 1.05$. **d** For $Jq/k_BT = 1.5$ (Interactive version at http://www.springer.com/cda/content/document/cda_downloaddocument/Selinger+Interactive+Figures.zip?SGWID=0-0-45-1509169-p177545420.)

Let us first consider the case where there is no applied field, $h = 0$. We begin at high temperature, so that Jq/k_BT is small. In that case, the free energy as a function of M has the form shown in Fig. 2.4a. It is approximately a parabola pointing upward, with the minimum at $M = 0$. Now we gradually reduce the temperature, i.e., increase the value of Jq/k_BT. As we reduce the temperature, the shape of the free energy around $M = 0$ gradually gets flatter and flatter. (I *strongly* recommend that you try the interactive figure to see this trend for yourself.) At a critical value $Jq/k_BT = 1$, shown in Fig. 2.4b, the minimum becomes so flat that the second derivative goes to zero. At this point, the curve no longer looks like a parabola; instead, it looks like a fourth-order function. As we reduce the temperature further, the single minimum at $M = 0$ splits up into two minima at small positive and negative values of M, as shown in Fig. 2.4c. As the temperature continues to decrease, the minima move outward and eventually approach ±1, as shown in Fig. 2.4d. (The cases with nonzero field in Fig. 2.5 will be discussed later.)

Notice what is happening here: At high temperature, the system goes to the state with no magnetic order, $M = 0$, as favored by entropy. By comparison, at low temperature, the system goes to a state with some magnetic order. Because there is no applied field, the system has no preference about whether the magnetic order should point up or down. Hence, it randomly chooses one of the minima, with positive or negative M. This low-temperature state with spontaneous (not induced by field) magnetic order is called a *ferromagnetic* phase. By contrast, the high-temperature

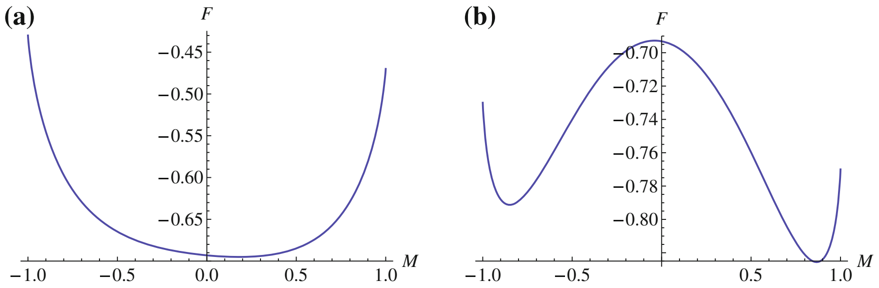


Fig. 2.5 Free energy of the *interacting* Ising model, with $h/k_B T = 0.02$, as a function of M. **a** For $Jq/k_B T = 0.9$. **b** For $Jq/k_B T = 1.5$ (Interactive version at http://www.springer.com/cda/content/document/cda_downloaddocument/Selinger+Interactive+Figures.zip?SGWID=0-0-45-1509169-p177545420.)

state with no spontaneous magnetic order is called a *paramagnetic* phase. The change from paramagnetic to ferromagnetic at a specific temperature is a *phase transition*.

Notice also a point of symmetry: The high-temperature paramagnetic phase has a symmetry between up and down, with no preference for either direction. In the low-temperature ferromagnetic phase, this symmetry is broken and the system randomly goes one way or the other. This random selection is called *spontaneous symmetry breaking*.

I should emphasize that the high-temperature state has *more symmetry* and it is *disordered*. By contrast, the low-temperature state has *less symmetry* and it is *more ordered*. In this sense, order means the opposite of symmetry; it means broken symmetry. (Students sometimes get confused about this point. They think that order is good and symmetry is good, so order must be the same thing as symmetry. No, it is not a question of good and bad!)

We might want to know the phase transition temperature precisely (not just by playing with the graphs). One feature of the phase transition that we already noticed is that the second derivative $\partial^2 F/\partial M^2 = 0$ at $M = 0$. Hence, we calculate the second derivative:

$$\frac{\partial^2}{\partial M^2}\left(\frac{F}{Nk_B T}\right) = -\frac{Jq}{k_B T} + \frac{1}{2(1+M)} + \frac{1}{2(1-M)}. \tag{2.32}$$

At $M = 0$, it becomes

$$\frac{\partial^2}{\partial M^2}\left(\frac{F}{Nk_B T}\right)\bigg|_{M=0} = 1 - \frac{Jq}{k_B T}. \tag{2.33}$$

Hence, the transition occurs at the temperature

$$T_C = \frac{Jq}{k_B}. \tag{2.34}$$

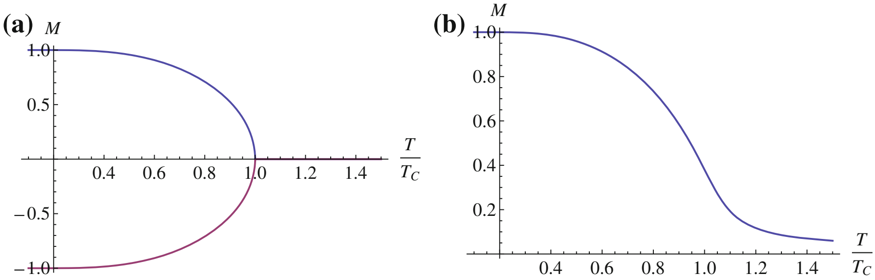


Fig. 2.6 Ising order parameter M as a function of temperature T. **a** For $h/k_BT = 0$. **b** For $h/k_BT = 0.02$ (Interactive version at http://www.springer.com/cda/content/document/cda_downloaddocument/Selinger+Interactive+Figures.zip?SGWID=0-0-45-1509169-p177545420.)

This temperature is called the *critical temperature*, and hence is given the notation T_C. Note that T_C is proportional to the interaction strength J and the coordination number q. If those quantities increase, then it is easier to get an ordered phase, so the phase transition occurs at a higher temperature.

We might also want to know how the equilibrium value of M varies with temperature, as the system cools from $T = T_C$ down to $T = 0$. One approach is to minimize the free energy numerically at each temperature. A plot of the numerical result is shown in Fig. 2.6a. Alternatively, we can try to find the minimum by setting the first derivative equal to zero:

$$\frac{\partial}{\partial M}\left(\frac{F}{Nk_BT}\right) = -\left(\frac{Jq}{k_BT}\right)M - \frac{h}{k_BT} + \tanh^{-1} M = 0. \tag{2.35}$$

We cannot solve this equation analytically in general, but we can make a good approximation when M is small (which is valid for T slightly below T_C). In this case, we approximate the inverse hyperbolic tangent by its Taylor series to obtain

$$\frac{\partial}{\partial M}\left(\frac{F}{Nk_BT}\right) \approx -\left(\frac{T_C}{T}\right)M - \frac{h}{k_BT} + \left(M + \frac{1}{3}M^3\right) = 0. \tag{2.36}$$

(In the first term of this expression, T_C is substituted in place of Jq/k_B.) For $h = 0$, the solution is

$$M = 0 \text{ or } M \approx \pm\sqrt{\frac{3(T_C - T)}{T}} \approx \pm\sqrt{\frac{3(T_C - T)}{T_C}}. \tag{2.37}$$

Note that $M = 0$ is the free energy minimum for $T > T_C$, and the free energy *maximum* for $T < T_C$. The nonzero solution is only defined for $T < T_C$, where it is the free energy minimum.

From Eq. (2.37), or just from looking at Fig. 2.6a, we see that M increases as the square root of the temperature difference when T drops slightly below T_C. This behavior is an example of a *scaling relation*. In general, people say that

$$M \propto (T_C - T)^{\beta}, \tag{2.38}$$

where the exponent β is called a *critical exponent*. Here, we see that $\beta = \frac{1}{2}$ in mean-field theory for the Ising model. This mean-field prediction is not exactly correct: More precise theories (and experiments!) show that $\beta = \frac{1}{8}$ for a 2D Ising system and $\beta \approx 0.31$ for a 3D Ising system. The mean-field prediction for the exponent is only exact for 4D or higher. However, mean-field theory is still teaching us a lot: It shows that there is a phase transition and that the order parameter has singular behavior just below the transition, which can be described by a power law.

So far we have only considered the interacting system when there is no applied field, but we can also consider the case with a field. If we apply a magnetic field h, it breaks the symmetry between $M > 0$ and $M < 0$, so the free energy is no longer an even function of M. For $T > T_C$, the free energy has the form shown in Fig. 2.5a. It has only a single minimum, and that minimum is not exactly at $M = 0$; it is displaced from zero by the applied field. By contrast, for $T < T_C$, the free energy has the form shown in Fig. 2.5b. It has two minima, and these minima are not equally deep. The applied field breaks the symmetry between the minima, and favors the state with magnetic order aligned with the field. Under an applied field, there is a smooth crossover from $T > T_C$ to $T < T_C$. Figure 2.6b shows the equilibrium value of M as a function of T for a small but nonzero field. Note that there is no singularity in M, i.e., no phase transition! Rather, M increases rapidly but smoothly around $T \approx T_C$. We can understand this smooth behavior by saying that the symmetry between $M > 0$ and $M < 0$ is already pre-broken by the field, so there is no need for a symmetry-breaking transition.

For an alternative view of the same physical behavior, imagine an experiment that varies the field at fixed temperature. For $T > T_C$, the behavior is shown in Fig. 2.7a. At this high temperature, there is no phase transition; rather M scales linearly with h for small fields, and then saturates for large fields. As T approaches T_C, the linear response to small field becomes sharper and sharper. When $T = T_C$, the linear response to small field becomes infinitely sharp, i.e., the plot of $M(h)$ has an infinite slope right at $h = 0$, as shown in Fig. 2.7b. When $T < T_C$, the response actually has a discontinuity at $h = 0$, as in Fig. 2.7c. Think about the free energy curve at this low temperature, which has two minima. As h passes through 0, the order parameter jumps from the minimum at $M < 0$ to the minimum at $M > 0$. The magnitude of this discontinuity increases as T decreases further below T_C, as in Fig. 2.7d. (You should compare the interactive Figs. 2.5 and 2.7 to see this behavior for yourself.)

Figure 2.8 shows a 3D plot of M as a function of both temperature T and field h. It provides the same information as in Figs. 2.6 and 2.7, but in a somewhat more beautiful visualization. This 3D plot looks like a partially torn sheet of paper. The

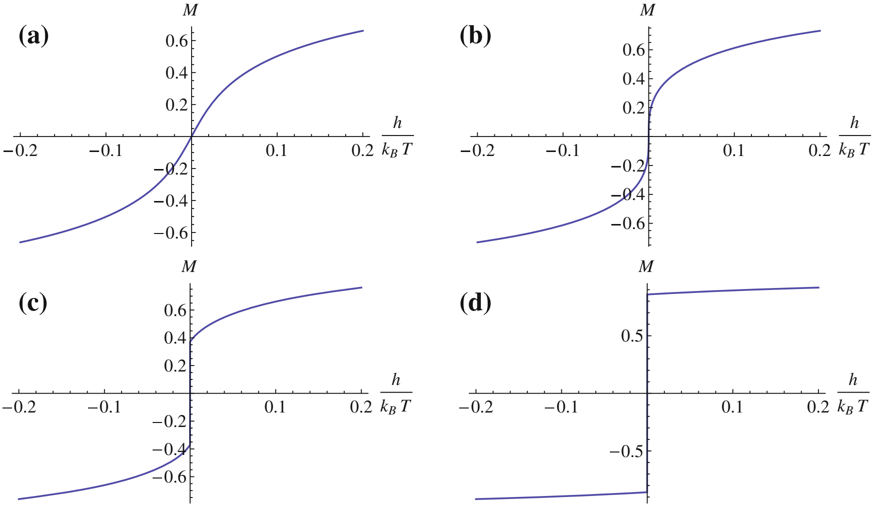


Fig. 2.7 Ising order parameter M as a function of field h. **a** For $Jq/k_BT = 0.9$. **b** For $Jq/k_BT = 1$. **c** For $Jq/k_BT = 1.05$. **d** For $Jq/k_BT = 1.5$ (Interactive version at http://www.springer.com/cda/content/document/cda_downloaddocument/Selinger+Interactive+Figures.zip?SGWID=0-0-45-1509169-p177545420.)

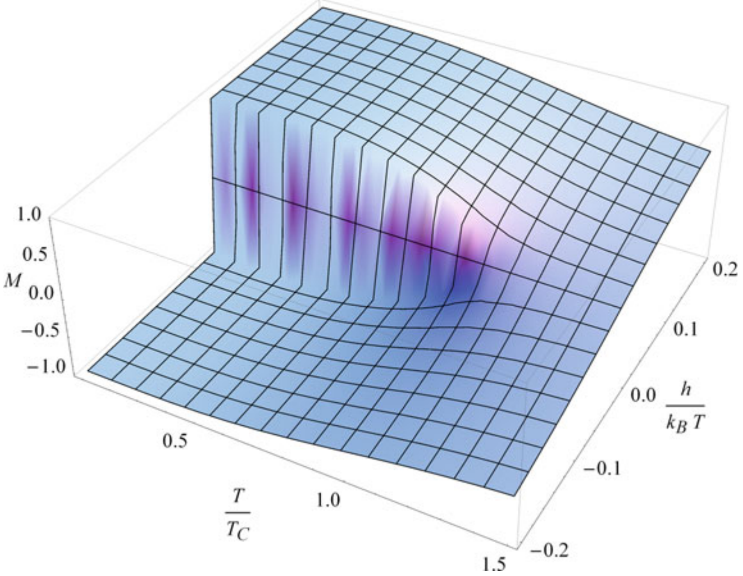


Fig. 2.8 Ising order parameter M as a function of temperature T and field h, in a 3D plot (Interactive version at http://www.springer.com/cda/content/document/cda_downloaddocument/Selinger+Interactive+Figures.zip?SGWID=0-0-45-1509169-p177545420.)

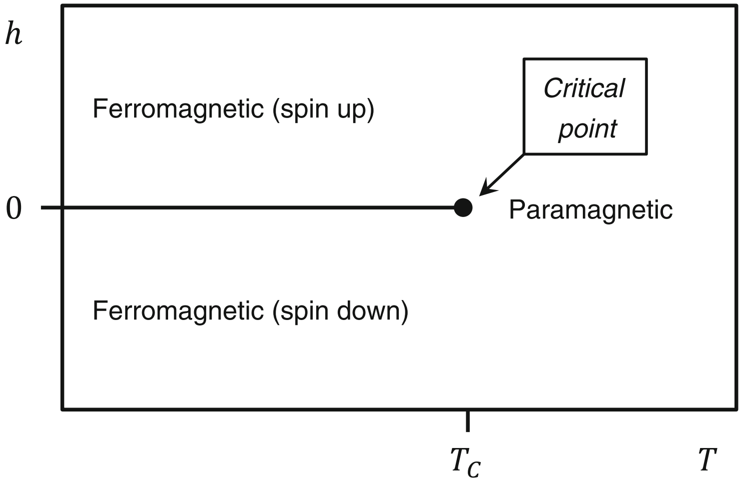


Fig. 2.9 Phase diagram of the Ising model as a function of temperature T and field h

"torn" part on the left represents the discontinuous response to a field for $T < T_C$. The "non-torn" part on the right represents the linear response to a field for $T > T_C$.

If we project this 3D plot down into the (T, h) plane, we obtain the *phase diagram* shown in Fig. 2.9. This phase diagram shows three types of behavior:

- For $T < T_C$, the system is ferromagnetic. Across the line $h = 0$, there is a *first-order phase transition* between spin-up and spin-down, with a discontinuous change in the order parameter M. This discontinuity corresponds to the "torn" part of the 3D plot. We can recognize the first-order transition in the free energy plots because there are two competing minima, and the system jumps from one minimum to the other.
- The point $T = T_C$ and $h = 0$, where the first-order transition terminates, is a very special point called the *critical point*, where the system has a second-order phase transition. At the second-order phase transition, there is no discontinuity in the order parameter, but there is a discontinuity in the derivative of the order parameter $\partial M/\partial T$. We can recognize the second-order transition in the free energy plots because the single minimum at high temperature becomes flat and breaks up into two minima at low temperature. At the critical point, the system has interesting singular behavior, which is characterized by critical exponents—see the discussion of the exponent β in Eq. (2.38), as well as the problems below.
- For $T > T_C$, the system is paramagnetic. As h varies between positive and negative values, there is no phase transition! There is just a disordered phase that responds smoothly to the applied field.

Problem: Calculate the response of the system to a small applied field in the paramagnetic phase, for $T > T_C$.

Solution: We return to Eq. 2.36 for the Ising order parameter M. If the applied field is small, then the induced M must also be small. In that case, we can neglect the M^3 term in comparison with the M terms in the equation, and we obtain

$$-\left(\frac{T_C}{T}\right)M - \frac{h}{k_B T} + M = 0. \tag{2.39}$$

The solution of this equation is

$$M = \frac{h}{k_B(T - T_C)}. \tag{2.40}$$

Hence, the susceptibility to an applied field is

$$\chi \equiv \left.\frac{\partial M}{\partial h}\right|_{h=0} = \frac{1}{k_B(T - T_C)}. \tag{2.41}$$

This result shows that the susceptibility diverges as $T \rightarrow T_C$, i.e., the system becomes more and more sensitive to an applied field as it approaches the critical point. The behavior is characterized by the scaling relation

$$\chi \propto (T - T_C)^{-\gamma}, \tag{2.42}$$

where γ is a critical exponent. Thus, we see that $\gamma = 1$ in mean-field theory. (Remember that Eq. (2.12) was derived for the non-interacting Ising model; here, we are generalizing it to the Ising model with interactions.)

Problem: Calculate the response of the system to a small applied field at the critical point, for $T = T_C$.

Solution: Again, we return to Eq. 2.36 for M. Right at $T = T_C$, the M terms in the equation cancel each other, so we cannot neglect the M^3 term. In that case, the equation becomes

$$-\frac{h}{k_B T} + \frac{1}{3}M^3 = 0, \tag{2.43}$$

and hence

$$M = \left(\frac{3h}{k_B T}\right)^{1/3}. \tag{2.44}$$

This result shows that the system has an infinite susceptibility at $T = T_C$, i.e., a nonlinear response to a small applied field, as shown in Fig. 2.7b. This behavior is characterized by the scaling relation

$$M \propto h^{1/\delta}, \tag{2.45}$$

where δ is another critical exponent. Thus, we see that $\delta = 3$ in mean-field theory.

2.3.2 Mean-Field Theory in Terms of Average Field

At this point, dear students, you have learned a lot about mean-field theory. You might be wondering, “Why is it called mean-field theory? I don’t see anything about mean fields in the theory!”

In this brief section, I will show you an alternative derivation of the theory, which shows the significance of the name. Some of you may have already seen this derivation in classes on magnetism, solid-state physics, or thermal physics—here I will show that it is equivalent to our previous derivation.

Let us go back to the energy of the Ising model defined in Eq. (2.1),

$$E = -J \sum_{\langle i,j \rangle} \sigma_i \sigma_j - h \sum_i \sigma_i . \tag{2.46}$$

We will consider a particular spin σ_i as our “test spin.” How does the energy depend on σ_i? Well, let us rewrite the energy as

$$E = -\left[h + J \left(\sum_{j=\text{neighbor of } i} \sigma_j \right) \right] \sigma_i + (\text{terms that do not depend on } \sigma_i). \tag{2.47}$$

This expression shows that our test spin σ_i experiences an effective field consisting of the actual field h plus an interaction with each of the neighboring spins. Of course, we do not know these neighboring spins; they may be correlated with the test spin in a complicated way. As our approximation, we neglect these correlations and replace each of these spins by its average value. Because the spins are all identical, the average value of each neighboring spin is $\langle \sigma_j \rangle = M$. Recall that the coordination number of the lattice is q, so that there are q neighbors in the sum. Hence, our approximation for the energy is

$$E = -h_{\text{mean}} \sigma_i + (\text{terms that do not depend on } \sigma_i), \tag{2.48}$$

where

$$h_{\text{mean}} = h + JqM. \tag{2.49}$$

In this expression, h_{mean} can be considered the effective field or average field or “mean” field acting on σ_i. (Now you see the terminology!)

From our discussion of the non-interacting Ising model, we derived Eq. (2.11) for the response to a field. Now we assume that σ_i has the same response to the effective field,

$$\langle \sigma_i \rangle = \tanh \left(\frac{h_{\text{mean}}}{k_B T} \right) . \tag{2.50}$$

Of course, our test spin σ_i is really just the same as all the other spins, so it must have the same expectation value $\langle \sigma_j \rangle = M$. Hence, we can combine Eqs. (2.49) and (2.50) to obtain

$$M = \tanh\left(\frac{h + JqM}{k_B T}\right). \qquad (2.51)$$

This equation represents the condition for *self-consistency: M* must satisfy this equation so that the effective field of the neighboring spins is consistent with the magnetic order of the test spin.

In the previous section, Eq. (2.35), we minimized the free energy by setting its first derivative equal to zero:

$$\frac{\partial}{\partial M}\left(\frac{F}{N k_B T}\right) = -\left(\frac{Jq}{k_B T}\right) M - \frac{h}{k_B T} + \tanh^{-1} M = 0. \qquad (2.52)$$

Note that this equation is exactly equivalent to the self-consistency equation! Hence, minimizing the free energy means the same thing as solving the self-consistency equation. All of our results from the previous section carry over unchanged.

In general, I prefer the approach based on free energy because it provides some extra information, compared with the approach based on self-consistency. At low temperature, the self-consistency equation has three solutions. By looking at the free energy plot, as in Figs. 2.4c, d and 2.5b, we can see that these solutions correspond to three places where $\partial F/\partial M = 0$: first a minimum, then a maximum, and then another minimum. We can disregard the maximum, and identify which of the local minima is the lowest, i.e., the *absolute minimum* of the free energy. If we did not have the free energy, we would have to fall back onto other physical arguments to select which solution of the self-consistency equation is correct.

Further Reading

The Ising model is discussed in many textbooks on magnetism, solid-state physics, and thermal physics, such as the books listed at the end of Chap. 1. For a discussion more advanced than the current chapter, I recommend:

1. P.M. Chaikin, T.C. Lubensky, *Principles of Condensed Matter Physics* (Cambridge, 1995)

If you are interested in the history of the Ising model, you might want to read Ising's original article:

2. E. Ising, Beitrag zur Theorie des Ferromagnetismus. Z. Phys. **31**, 253–258 (1925). doi:10.1007%2FBF02980577

as well as the historical review article:

3. S.G. Brush, History of the Lenz-Ising Model. Rev. Mod. Phys. **39**, 883–893 (1967). doi:10.1103/RevModPhys.39.883

Chapter 3
Gases and Liquids

Abstract This chapter extends the previous discussion of the Ising model to a system of molecules forming gas or liquid phases. It begins with the ideal gas of non-interacting molecules and derives the ideal gas law by minimizing the Gibbs free energy. It then moves on to the statistical mechanics of interactingmolecules—a much more complex problem that cannot be solved exactly. A useful approximation is van der Waals theory, which is a type of mean-field theory, analogous to mean-field theory for the interacting Ising model. Van der Waals theory shows that the interacting system has distinct gas and liquid phases, and makes predictions for the phase diagram.

I would now like to move on to a different type of statistical mechanics problem: a system of molecules forming a gas or liquid phase. This problem is more interesting than the Ising model because it is a molecular system, so it takes us on our way toward describing liquid crystals and other complex molecular systems. We will now have to keep track of the positions and momenta of molecules, not just spins on a lattice. In spite of this difference, we will still be able to use the concepts of energy, entropy, and free energy to describe the behavior of the gas or liquid, and we will see that the phase diagram is actually related to that of the Ising model.

3.1 Ideal Gas at Fixed Volume

As a first step in this problem, let us consider an "ideal gas." This term means that the gas molecules do not interact with each other; they just move around freely, confined by their container, as shown in Fig. 3.1. That approximation is generally reasonable when the gas molecules are very dilute, i.e., the number of molecules per unit volume is very low.

Following the procedure from the previous chapters, we want to calculate the statistical properties of the ideal gas by averaging over all the states of the system. So we have to ask: What is a state? For the Ising model, a state means a configuration of all the spins σ_1, σ_2, ... When we sum over the states, we sum over both possibilities for σ_1, both possibilities for σ_2, etc. For the ideal gas, each molecule i is characterized by its position $\boldsymbol{r}_i$ and its momentum $\boldsymbol{p}_i$. For each molecule, we

J.V. Selinger, *Introduction to the Theory of Soft Matter*,
Soft and Biological Matter, DOI 10.1007/978-3-319-21054-4_3

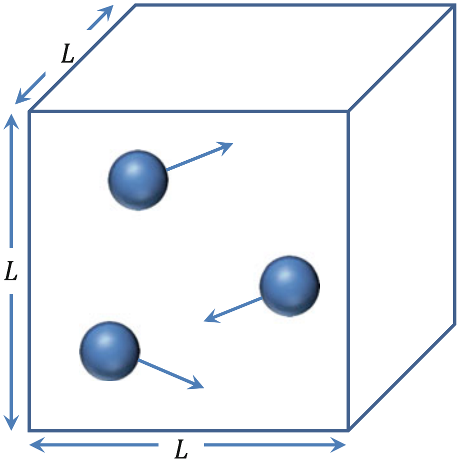


Fig. 3.1 Dilute system of molecules moving inside a box of size $L \times L \times L$

need to specify three vector components of position and three vector components of momentum, or six numbers in all. Hence, if there are N molecules in the system, we need to specify $6N$ numbers. People often refer to the $6N$-dimensional space given by $\{\boldsymbol{r}_1, \boldsymbol{p}_1, \boldsymbol{r}_2, \boldsymbol{p}_2, \ldots \boldsymbol{r}_N, \boldsymbol{p}_N\}$ as *phase space*. Each state of the system is thus given by a point in phase space.

For every state of the system, there is a certain energy. For the moment, let us suppose that there is no potential energy; there is only kinetic energy. In this case, the energy of a state is just

$$E(\boldsymbol{r}_1, \boldsymbol{p}_1, \boldsymbol{r}_2, \boldsymbol{p}_2, \ldots \boldsymbol{r}_N, \boldsymbol{p}_N) = E_{\text{kinetic}}(\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots \boldsymbol{p}_N), \tag{3.1}$$

where

$$E_{\text{kinetic}}(\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots \boldsymbol{p}_N) = \sum_{i=1}^{N} \frac{|\boldsymbol{p}_i|^2}{2m}. \tag{3.2}$$

Here, m is the mass of each molecule. Note that the kinetic energy depends on the momenta but not on the positions.

The partition function Z is now the sum of e^{-E/k_BT} over all the states of the system. Because the positions and momenta are continuous variables, we cannot just sum over a discrete set of possibilities; instead, we must *integrate* over these variables. Hence, we might expect the partition function to be given by the $6N$-dimensional integral:

$$Z \overset{?}{=} \int d^3\boldsymbol{r}_1 d^3\boldsymbol{p}_1 d^3\boldsymbol{r}_2 d^3\boldsymbol{p}_2 \cdots d^3\boldsymbol{r}_N d^3\boldsymbol{p}_N e^{-E(\boldsymbol{r}_1,\boldsymbol{p}_1,\boldsymbol{r}_2,\boldsymbol{p}_2,\ldots\boldsymbol{r}_N,\boldsymbol{p}_N)/k_BT}. \tag{3.3}$$

This expression is *almost* correct. We just need to fix two small problems:

The first problem is that the units of Z are wrong. The partition function is defined as the sum over states of e^{-E/k_BT}. This exponential is a dimensionless quantity, and hence the sum of many exponential terms must also be a dimensionless quantity. However, in the expression of Eq. (3.3), each integral over position gives units of (length)^3, and each integral over momentum gives units of (momentum)^3. In other words, this expression for the partition function has units of (length · momentum)^{3N}, but it should be dimensionless. Hence, we need to divide by something with units of (length · momentum), raised to the $3N$ power. From your quantum mechanics classes, you may recall that one fundamental physical constant has units of (length · momentum): Planck's constant h. Hence, to fix the units, we ought to divide the $6N$-dimensional integral by a factor of h^{3N}.

There is actually an important physical reason why Planck's constant comes into the partition function. Earlier I said that each physical region corresponds to a point in phase space, as if the state has an exact position and an exact momentum for each molecule. However, from Heisenberg's Uncertainty Principle, you may recall that we cannot exactly specify the position and momentum of a particle. Rather, there is always some uncertainty in the position and momentum, which is given by $\Delta x \Delta p_x \approx h$. Thus, a state cannot really be a precise point in phase space; rather, it must be a fuzzy region in phase space. The "effective volume" of the fuzzy region is h for each dimension (x, y, or z) for each molecule (1 through N), and hence h^{3N}. When we sum over states, we must integrate over all phase space and divide by the "effective volume" h^{3N} of each state in this $6N$-dimensional dimensional space. In this sense, our theory has a hidden quantum mechanics aspect. This quantum mechanics will not really affect any of our results, but it is needed to get the right units.

The second problem is that the particles are indistinguishable. For that reason, it does not matter whether molecule 1 is at a certain position and momentum and molecule 2 is at another position and momentum, or vice versa; these are the same physical state. In other words, we can relabel the molecules in any arbitrary way, and still have the same state. The number of ways to relabel the N particles is $N!$. Hence, to account for indistinguishability, we ought to divide the $6N$-dimensional integral by a factor of $N!$.

With these two corrections, the partition function for the ideal gas becomes

$$\begin{aligned} Z &= \frac{1}{N!h^{3N}} \int d^3\boldsymbol{r}_1 d^3\boldsymbol{p}_1 d^3\boldsymbol{r}_2 d^3\boldsymbol{p}_2 \cdots d^3\boldsymbol{r}_N d^3\boldsymbol{p}_N e^{-E(\boldsymbol{r}_1,\boldsymbol{p}_1,\boldsymbol{r}_2,\boldsymbol{p}_2,\ldots\boldsymbol{r}_N,\boldsymbol{p}_N)/k_BT} \\ &= \frac{1}{N!h^{3N}} \int d^3\boldsymbol{r}_1 d^3\boldsymbol{p}_1 d^3\boldsymbol{r}_2 d^3\boldsymbol{p}_2 \cdots d^3\boldsymbol{r}_N d^3\boldsymbol{p}_N \exp\left[-\sum_{i=1}^{N} \frac{|\boldsymbol{p}_i|^2}{2mk_BT}\right]. \end{aligned} \tag{3.4}$$

Because there are no interactions between different molecules, this partition function factorizes as

$$Z = \frac{1}{N!}\left[\frac{1}{h^3}\int d^3\boldsymbol{r} d^3\boldsymbol{p} \exp\left(-\frac{|\boldsymbol{p}|^2}{2mk_BT}\right)\right]^N. \tag{3.5}$$

Inside the square brackets, the integral over position gives a factor of volume $V = L^3$. The integral over momentum is a Gaussian integral, which can be done exactly. The result is

$$Z = \frac{1}{N!}\left[\frac{V(2\pi m k_B T)^{3/2}}{h^3}\right]^N . \tag{3.6}$$

Note that this problem is analogous to the *non-interacting* Ising model, because the lack of interactions allows the partition function to factorize and hence to be calculated exactly.

Now that we have the partition function, we can find the free energy

$$\begin{aligned} F =& - k_B T \log Z \qquad (3.7)\\ =& - N k_B T \log\left[\frac{V(2\pi m k_B T)^{3/2}}{h^3}\right] + k_B T \log N!. \end{aligned}$$

Using Stirling's formula $\log N! \approx N \log N - N$, it reduces to

$$F = -N k_B T \log\left[\frac{V(2\pi m k_B T)^{3/2} e}{N h^3}\right]. \tag{3.8}$$

Note that the free energy is *extensive:* if we double the number of molecules *and* we double the volume, then we get double the free energy, as we should. (If we double the number of molecules and we do not change the volume, then we have a different *denser* system, and the free energy changes in a more complicated way.)

Now that we have the partition function and the free energy, we can calculate many other statistical properties. The probability of any state is

$$\text{Prob}(\boldsymbol{r}_1, \boldsymbol{p}_1, \ldots \boldsymbol{r}_N, \boldsymbol{p}_N) = \frac{1}{Z}\exp\left[-\sum_{i=1}^{N}\frac{|\boldsymbol{p}_i|^2}{2mk_B T}\right]. \tag{3.9}$$

The probabilities are correctly normalized, because the sum over all states of the probabilities is 1:

$$1 = \frac{1}{N! h^{3N}}\int d^3\boldsymbol{r}_1 d^3\boldsymbol{p}_1 \cdots d^3\boldsymbol{r}_N d^3\boldsymbol{p}_N \frac{1}{Z}\exp\left[-\sum_{i=1}^{N}\frac{|\boldsymbol{p}_i|^2}{2mk_B T}\right], \tag{3.10}$$

by the definition of the partition function. Hence, the expectation value of any physical quantity X is

$$\langle X\rangle = \frac{1}{N! h^{3N}}\int d^3\boldsymbol{r}_1 d^3\boldsymbol{p}_1 \cdots d^3\boldsymbol{r}_N d^3\boldsymbol{p}_N (X)\frac{1}{Z}\exp\left[-\sum_{i=1}^{N}\frac{|\boldsymbol{p}_i|^2}{2mk_B T}\right]. \tag{3.11}$$

For example, the average kinetic energy of molecule 1 is

$$\left\langle \frac{|\boldsymbol{p}_1|^2}{2m} \right\rangle = \frac{1}{N!h^{3N}} \int d^3\boldsymbol{r}_1 d^3\boldsymbol{p}_1 \cdots d^3\boldsymbol{r}_N d^3\boldsymbol{p}_N \left[\frac{|\boldsymbol{p}_1|^2}{2m} \right] \frac{1}{Z} \exp\left[-\sum_{i=1}^{N} \frac{|\boldsymbol{p}_i|^2}{2mk_B T} \right]. \tag{3.12}$$

Because of the factor of Z in the denominator, almost all of the integrals cancel, and we are left with only

$$\left\langle \frac{|\boldsymbol{p}_1|^2}{2m} \right\rangle = \frac{\displaystyle\int d^3\boldsymbol{p}_1 \left[\frac{|\boldsymbol{p}_1|^2}{2m} \right] \exp\left[-\frac{|\boldsymbol{p}_1|^2}{2mk_B T} \right]}{\displaystyle\int d^3\boldsymbol{p}_1 \exp\left[-\frac{|\boldsymbol{p}_1|^2}{2mk_B T} \right]} = \frac{3}{2} k_B T, \tag{3.13}$$

calculating the Gaussian integrals exactly.[1] Of course, all of the molecules are identical, so they all have the same average kinetic energy. Hence, the average kinetic energy for all N molecules is

$$\langle E_{\text{kinetic}} \rangle = \frac{3}{2} N k_B T. \tag{3.14}$$

As you recall, in this calculation, we are assuming that there is no potential energy. For that reason, the average total energy is the same as the average kinetic energy:

$$\langle E \rangle = \frac{3}{2} N k_B T. \tag{3.15}$$

The average energy $\langle E \rangle$ is part of the free energy $F = \langle E \rangle - TS$. By comparing our expressions for the free energy (3.8) and the average energy (3.15), we can obtain the entropy

$$S = \frac{\langle E \rangle - F}{T} = N k_B \log \left[\frac{V(2\pi m k_B T)^{3/2} e^{5/2}}{N h^3} \right]. \tag{3.16}$$

3.2 Ideal Gas at Fixed Pressure

In the previous section, we considered an ideal gas at fixed volume. Now suppose that it is free to change its volume—for example, it might be in a container that is free to expand or contract. What volume will it select?

First of all, let us discuss the states of this system with variable volume. The volume is now one extra degree of freedom, in addition to the positions and momenta of all the molecules. We might say that the phase space is $(6N + 1)$-dimensional,

[1] This result is a special case of the *equipartition theorem:* Any degree of freedom that enters *quadratically* in the energy function gets $\frac{1}{2} k_B T$ of energy in thermal equilibrium. The kinetic energy of particle 1 contains three such degrees of freedom (p_{1x}, p_{1y}, and p_{1z}).

including this extra degree of freedom. Each microstate of the system is characterized by all $(6N + 1)$ degrees of freedom—the positions, momenta, and volume. When we include the variable volume, the partition function becomes

$$Z_{\text{var vol}} = \int \frac{dV}{V_0} \underbrace{\frac{1}{N!h^{3N}} \int d^3\boldsymbol{r}_1 d^3\boldsymbol{p}_1 \cdots d^3\boldsymbol{r}_N d^3\boldsymbol{p}_N e^{-E(\boldsymbol{r}_1,\boldsymbol{p}_1,\boldsymbol{r}_2,\boldsymbol{p}_2,\ldots\boldsymbol{r}_N,\boldsymbol{p}_N)/k_BT}}_{e^{-F(V)/k_BT}}, \tag{3.17}$$

where V_0 is a constant with units of volume to make Z dimensionless.

I will now classify the microstates into macrostates *based on the volume*. When I do this classification, each macrostate has a free energy $F(V)$, which can be calculated by summing over all the microstates with that volume V. Luckily, we already did that calculation in the previous section! The result is given in Eq. (3.8):

$$F(V) = -Nk_BT \log\left[\frac{V(2\pi mk_BT)^{3/2}e}{Nh^3}\right]. \tag{3.18}$$

We can now use this free energy to compare macrostates with different volumes. Note that this calculation is analogous to the Ising model in the previous chapter. There I classified microstates into macrostates based on the magnetic order parameter M. Hence, we might say that V in this problem is analogous to M in that problem.

Now let us consider how the free energy $F(V)$ depends on volume. We can see that it is proportional to $-\log(\text{const} \cdot V)$. Hence, as the volume increases, the free energy decreases monotonically; there is no minimum. Hence, we might expect that the ideal gas will expand forever, with $V \to \infty$. This result is disturbing; we were expecting an equilibrium volume.

The problem is that we forgot about *pressure!* If we release an ideal gas in outer space, where there is no pressure, then it will indeed expand to infinity. However, if we do the experiment on Earth, under conditions of ambient pressure, the situation is different: Whenever the ideal gas expands, it must push away something else. Expanding against a pressure p requires an energy cost of $+pV$. When we include this energy cost, we transform the free energy into the *Gibbs free energy*:

$$G = F + pV. \tag{3.19}$$

Hence, for the ideal gas, the Gibbs free energy becomes

$$G(V) = -Nk_BT \log\left[\frac{V(2\pi mk_BT)^{3/2}e}{Nh^3}\right] + pV. \tag{3.20}$$

Note that the first term of $G(V)$ decreases and the second term increases as V increases. Hence, $G(V)$ can indeed have a minimum, corresponding to an equilibrium volume.

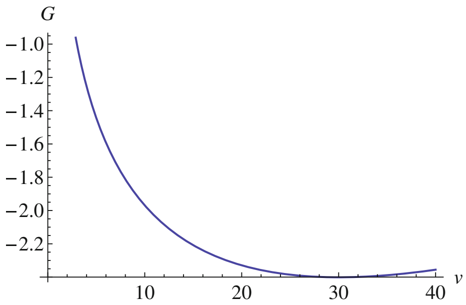


Fig. 3.2 Gibbs free energy of an ideal gas, as a function of the volume per molecule $v = V/N$. Parameters for this plot are $p = 0.01$ and $k_BT = 0.3$ (Interactive version at http://www.springer.com/cda/content/document/cda_downloaddocument/Selinger+Interactive+Figures.zip?SGWID=0-0-45-1509169-p177545420.)

Figure 3.2 shows a plot of the Gibbs free energy as a function of volume. For this plot, it is convenient to normalize $G(V)$ by a factor of Nk_BT, and plots the result as a function of the volume per molecule $v = V/N$. Our function then becomes

$$\frac{G(v)}{Nk_BT} = -\log\left[\left(\frac{(2\pi mk_BT)^{3/2}e}{h^3}\right)v\right] + \left(\frac{p}{k_BT}\right)v. \tag{3.21}$$

In the figure, we can see that $G(v)$ has a single minimum. In the interactive version of the figure, we can shift the pressure and temperature. The equilibrium value of v moves to smaller volume as pressure increases, and to larger volume as temperature increases.

To derive an algebraic expression for equilibrium volume, we can just minimize the Gibbs free energy by setting its first derivative with respect to volume equal to zero:

$$\frac{\partial G}{\partial V} = 0. \tag{3.22}$$

This minimization gives the relationship between pressure and volume, sometimes called the *equation of state:*

$$pV = Nk_BT. \tag{3.23}$$

You should recognize this expression as the ideal gas law!

Incidentally, some students are more familiar with the ideal gas law in the form

$$pV = nRT, \tag{3.24}$$

where n is the number of moles and R is the gas constant. These two expressions are exactly equivalent. To see the equivalence, just multiply and divide Eq. (3.23) by Avogadro's number $N_A = 6 \times 10^{23}$ to obtain

$$pV = \underbrace{\frac{N}{6 \times 10^{23}}}_{n} \underbrace{(6 \times 10^{23} k_B)}_{R} T. \tag{3.25}$$

Here, the number of moles n is related to the number of molecules N by $n = N/(6 \times 10^{23})$, and the gas constant R is related to Boltzmann's constant k_B by $R = 6 \times 10^{23} k_B$, because R involves energy per mole rather than energy per molecule.

3.3 Ideal Gas Under Gravity

For one more variation on the ideal gas, let us consider an ideal gas in a gravitational field. This problem is analogous to the non-interacting Ising model in a magnetic field, because the gas molecules each interact with the gravitational field but they do not interact with each other.

Suppose we want to model a gas above a flat surface of size $L \times L$ in the xy-plane. In the vertical direction, the gas molecules can go anywhere from $z = 0$ to $z = \infty$. Because of the gravitational field, the system has both kinetic and potential energies, so the total energy of a state is

$$E(\boldsymbol{r}_1, \boldsymbol{p}_1, \boldsymbol{r}_2, \boldsymbol{p}_2, \ldots \boldsymbol{r}_N, \boldsymbol{p}_N) = \sum_{i=1}^{N} \left(\frac{|\boldsymbol{p}_i|^2}{2m} + mgz_i \right), \tag{3.26}$$

where g is the gravitational constant and z_i is the height of molecule i above the bottom surface. Hence, the partition function is

$$Z = \frac{1}{N! h^{3N}} \int d^3\boldsymbol{r}_1 d^3\boldsymbol{p}_1 \cdots d^3\boldsymbol{r}_N d^3\boldsymbol{p}_N \exp\left[-\sum_{i=1}^{N} \left(\frac{|\boldsymbol{p}_i|^2}{2mk_B T} + \frac{mgz_i}{k_B T} \right) \right]. \tag{3.27}$$

Because the molecules do not interact with each other, this partition function factorizes as

$$Z = \frac{1}{N!} \left[\frac{1}{h^3} \int d^3\boldsymbol{p} \exp\left(-\frac{|\boldsymbol{p}|^2}{2mk_B T} \right) \int_0^L dx \int_0^L dy \int_0^\infty dz \exp\left(-\frac{mgz}{k_B T} \right) \right]^N. \tag{3.28}$$

As in the previous case, the momentum integral gives a factor of $(2\pi mk_B T)^{3/2}$, and the x and y integrals each give a factor of L. The only new calculation is the z integral, which gives a factor of $k_B T/mg$. As a result, the partition function becomes

$$Z = \frac{1}{N!}\left[\frac{(2\pi m k_B T)^{3/2} L^2}{h^3}\frac{k_B T}{mg}\right]^N. \tag{3.29}$$

With that partition function, the probability of any state becomes

$$\text{Prob}(\boldsymbol{r}_1, \boldsymbol{p}_1, \ldots \boldsymbol{r}_N, \boldsymbol{p}_N) = \frac{1}{Z}\exp\left[-\sum_{i=1}^{N}\left(\frac{|\boldsymbol{p}_i|^2}{2mk_B T} + \frac{mgz_i}{k_B T}\right)\right]. \tag{3.30}$$

From the general probability distribution function, we might want to calculate the probability that a particular molecule 1 is at height z_1, regardless of all the other degrees of freedom. We can calculate it as

$$\text{Prob}(z_1) = \frac{1}{N!h^{3N}}\int dx_1 dy_1 d^3\boldsymbol{p}_1 d^3\boldsymbol{r}_2 \cdots d^3\boldsymbol{r}_N d^3\boldsymbol{p}_N \text{Prob}(\boldsymbol{r}_1, \boldsymbol{p}_1, \ldots \boldsymbol{r}_N, \boldsymbol{p}_N), \tag{3.31}$$

integrating over all variables *except* z_1. It reduces to

$$\text{Prob}(z_1) = \frac{mg}{k_B T}\exp\left(-\frac{mgz_1}{k_B T}\right), \tag{3.32}$$

which is the Boltzmann distribution for molecular height. Of course, it is the same for any molecule, so we can just write it as $\text{Prob}(z)$. We might also want to calculate the average height of a molecule,

$$\langle z \rangle = \int_0^\infty dz(z)\text{Prob}(z) = \frac{k_B T}{mg}. \tag{3.33}$$

This expression shows that increasing temperature causes the average height to be higher, while increasing the molecular mass or the gravitational field causes the average height to be lower.

Problem: Apply this theory to the Earth's atmosphere and compare your answer with the actual thickness of the atmosphere.

Solution: As a rough estimate, let us assume that the atmosphere is mainly composed of nitrogen N_2 molecules, which have a molecular weight of 28 amu $= 4.6 \times 10^{-26}$ kg. Furthermore, let us assume that the atmosphere is at a uniform temperature of 300 K. In that case, the average height of a molecule above the Earth's surface should be

$$\langle z \rangle = \frac{k_B T}{mg} = \frac{(1.4 \times 10^{-23}\ \text{J/K})(300\ \text{K})}{(4.6 \times 10^{-26}\ \text{kg})(9.8\ \text{m/s}^2)} = 9300\ \text{m} = 9.3\ \text{km}. \tag{3.34}$$

This estimate is similar to the thickness of the troposphere, which is about 11 km. (It is much less than what space scientists call the "thickness of the atmosphere," but that is because space scientists are not interested in the average height of a molecule;

they are interested in the altitude where the atmosphere is so thin that satellites can stay in orbit without substantial air drag.)

Problem: Calculate the expectation value $\langle z^2 \rangle$. Is $\langle z^2 \rangle$ equal to $\langle z \rangle^2$?
Solution: Following the same method as in Eq. (3.33), we obtain

$$\langle z^2 \rangle = \int_0^\infty dz (z^2) \mathrm{Prob}(z) = 2 \left(\frac{k_B T}{mg} \right)^2 . \tag{3.35}$$

Note that $\langle z^2 \rangle$ is *not* equal to $\langle z \rangle^2$. It is reasonable that these quantities should be different, because the molecules have a distribution of heights. The difference $\langle z^2 \rangle - \langle z \rangle^2$ shows the variance of the heights. These quantities would only be equal if the molecules had a single unique height.

3.4 Gas with Interactions: van der Waals Theory

Let us now move on to consider a more interesting and complex problem: a gas of molecules that interact with each other. We would like to understand how the interactions change the behavior that was derived in the previous sections.

In general, we might write the potential energy of interacting molecules as the sum of many pairwise interactions

$$E_{\text{potential}}(\boldsymbol{r}_1, \ldots \boldsymbol{r}_N) = \sum_{(i,j)} U(|\boldsymbol{r}_i - \boldsymbol{r}_j|), \tag{3.36}$$

where the sum is taken over all pairs of molecules (i, j). The interaction $U(|\boldsymbol{r}_i - \boldsymbol{r}_j|)$ is a function of the distance between the centers of molecules i and j. The exact form of the interaction will, of course, depend on what type of molecules we are studying; it will be different for nitrogen, oxygen, carbon dioxide, etc. Nevertheless, the interaction usually has the generic shape shown in Fig. 3.3. This interaction has two important features. First, at short distances, the potential energy becomes

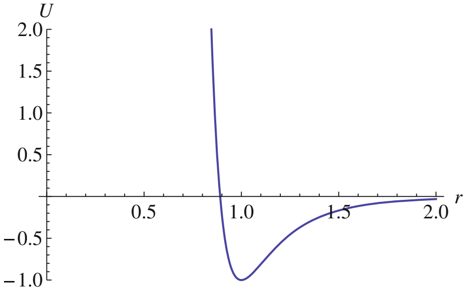


Fig. 3.3 Typical form for the interaction between two molecules, as a function of the distance between their centers. (The specific function plotted here is the Lennard-Jones potential $U(r) = Ar^{-12} + Br^{-6}$, which is a commonly used form of the interaction.)

extremely high. Because it is so high, the molecules effectively have hard cores, so they cannot overlap. Second, at longer distances, the potential becomes somewhat attractive. Hence, the molecules have some optimal separation away from each other.

We can try to put the potential energy into the partition function:

$$Z = \frac{1}{N!h^{3N}} \int d^3\boldsymbol{r}_1 d^3\boldsymbol{p}_1 \cdots d^3\boldsymbol{r}_N d^3\boldsymbol{p}_N \\ \exp\left[-\sum_{i=1}^{N} \frac{|\boldsymbol{p}_i|^2}{2mk_BT} - \sum_{(i,j)} \frac{U(|\boldsymbol{r}_i - \boldsymbol{r}_j|)}{k_BT} \right]. \qquad (3.37)$$

Unfortunately, we now have a serious problem. The momentum integrals still factorize nicely, but the position integrals are all coupled together and do not factorize into the product of independent integrals. This problem is essentially the same as the partition function for interacting Ising spins in Chap. 2. The position integral for molecule 1 is coupled to the position integral for molecule 2, which is coupled to the position integral for molecule 3, and so forth. Because the system has an enormous number of molecules, probably 10^{23} or more, we have an enormous number of integrals that are all coupled together. This is a profoundly difficult problem, and it cannot be solved exactly.

At this point, we need to make an approximation. As in the previous chapter, we will use *mean-field theory*. The particular version of mean-field theory applied to interacting gas molecules is called *van der Waals theory*. In van der Waals theory, we assume that each molecule moves in an effective environment composed of the the other molecules. We neglect the correlations between nearby molecules, and just think about the average properties of the environment. Figure 3.4 shows a schematic illustration (or "cartoon") of this concept. Here, one of the molecules is labeled in

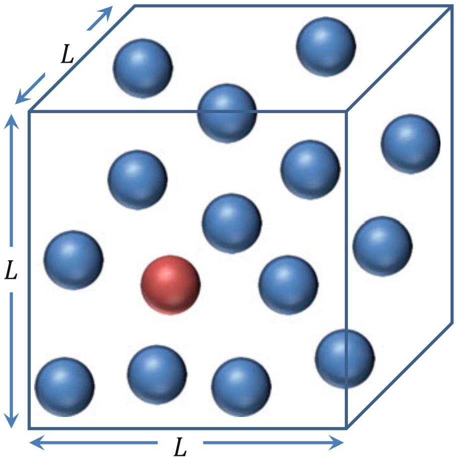


Fig. 3.4 Schematic illustration of mean-field theory for interacting gas molecules. One of the molecules is labeled in *red* as a "test" molecule, and it moves in an effective environment composed of *blue* molecules

red as our "test" molecule. It moves in an effective environment composed of the other molecules, which are blue. We want to calculate the statistical properties of the red molecule in that environment. Of course, all the blue molecules are identical to the red molecule, so they have the same statistical properties. Hence, we will need to find a self-consistent solution for the properties of the red and blue molecules.

To set up van der Waals theory mathematically, we suppose that each molecule moves in an effective potential $U_{\text{eff}}(\boldsymbol{r})$, which arises from the average interaction with the other molecules. In this case, the problem becomes similar to the ideal gas in a gravitational field, as discussed in the previous section. The partition function factorizes to give

$$Z = \frac{1}{N!}\left[\frac{1}{h^3}\int d^3\boldsymbol{p}\exp\left(-\frac{|\boldsymbol{p}|^2}{2mk_BT}\right)\int d^3\boldsymbol{r}\exp\left(-\frac{U_{\text{eff}}(\boldsymbol{r})}{k_BT}\right)\right]^N. \tag{3.38}$$

As usual, the momentum integral gives a factor of $(2\pi mk_BT)^{3/2}$, and we are left with

$$Z = \frac{1}{N!}\left[\frac{(2\pi mk_BT)^{3/2}}{h^3}\int d^3\boldsymbol{r}\exp\left(-\frac{U_{\text{eff}}(\boldsymbol{r})}{k_BT}\right)\right]^N. \tag{3.39}$$

Now what can we say about the position integral? First of all, certain regions of the volume are forbidden to the test molecule, i.e., $U_{\text{eff}}(\boldsymbol{r}) \to \infty$, because these regions are already occupied by other molecules. The *excluded volume* must be proportional to the number of other molecules, so we write it as bN, where b be a parameter with units of volume, which represents the excluded volume per molecule.[2] Hence, the integrand $e^{-U_{\text{eff}}(\boldsymbol{r})/k_BT} = 0$ in the excluded volume; it is only nonzero in the non-excluded volume $(V - bN)$.

Second, how big is $U_{\text{eff}}(\boldsymbol{r})$ in the non-excluded volume? Throughout that region, the test molecule experiences an effective negative potential arising from the other molecules. The magnitude of this effective negative potential must be proportional to the density of other molecules N/V. Hence, we write $U_{\text{eff}}(\boldsymbol{r}) = -aN/V$, where a is a parameter with units of energy·volume, which represents the strength of the negative potential in Fig. 3.3 multiplied by the interaction volume.

We now have a model with two parameters, b and a. In terms of these parameters, the position integral becomes $(V - bN)e^{aN/V}$, and hence the partition function becomes

$$Z = \frac{1}{N!}\left[\frac{(2\pi mk_BT)^{3/2}}{h^3}(V - bN)e^{aN/Vk_BT}\right]^N. \tag{3.40}$$

With this mean-field approximation for the partition function, we can calculate the free energy as

[2] Really, it is $b(N-1)$ for the molecules *other than* the test molecule, but $N \gg 1$, so the -1 is negligible.

$$\begin{aligned} F &= -\,k_B T \log Z \\ &= -\,N k_B T \log \left[\frac{(2\pi m k_B T)^{3/2} e}{h^3} \left(\frac{V}{N} - b \right) \right] - \frac{aN^2}{V}. \end{aligned} \tag{3.41}$$

At fixed pressure p, the corresponding Gibbs free energy is

$$\begin{aligned} G &= F + pV \\ &= -N k_B T \log \left[\frac{(2\pi m k_B T)^{3/2} e}{h^3} \left(\frac{V}{N} - b \right) \right] - \frac{aN^2}{V} + pV. \end{aligned} \tag{3.42}$$

As in the ideal gas case, we can minimize the Gibbs free energy to find the equilibrium volume:

$$\frac{\partial G}{\partial V} = 0, \tag{3.43}$$

which implies

$$\left(p + \frac{aN^2}{V^2} \right) (V - bN) = N k_B T. \tag{3.44}$$

You may recognize Eq. (3.44) as the *van der Waals equation of state*. Note that it reduces to the ideal gas law if $b \to 0$ and $a \to 0$.

One way to think about van der Waals theory is as a correction to the ideal gas theory, which takes intermolecular interactions into account. This perspective might be useful for chemical engineers, for example, if they need to know the relationship between pressure, volume, and temperature more precisely than they would get from the ideal gas law. They might look up the parameters b and a for oxygen or carbon dioxide, and hence calculate the right relationship between pressure, volume, and temperature for the gas that they are working with.

3.5 Gas–Liquid Transition at Fixed Pressure

We can learn much more from van der Waals theory than just these detailed corrections for specific materials. Let us plot the Gibbs free energy over a range of parameters. In these plots, we can scale G by $N k_B T$, and express the result in terms of the volume per molecule $v = V/N$, to obtain

$$\frac{G}{N k_B T} = -\log \left[\frac{(2\pi m k_B T)^{3/2} e}{h^3} (v - b) \right] - \left(\frac{a}{k_B T} \right) \frac{1}{v} + \left(\frac{p}{k_B T} \right) v. \tag{3.45}$$

For the plots, we can choose units of volume such that $b = 1$, and units of energy such that $a = 1$. (Of course, whenever we want to apply our results to specific materials, we will have to go back to physical units!)

Figure 3.5 shows a sequence of plots as the temperature is reduced, at a fixed low pressure. In Fig. 3.5a, at high temperature, the plot looks very similar to the ideal

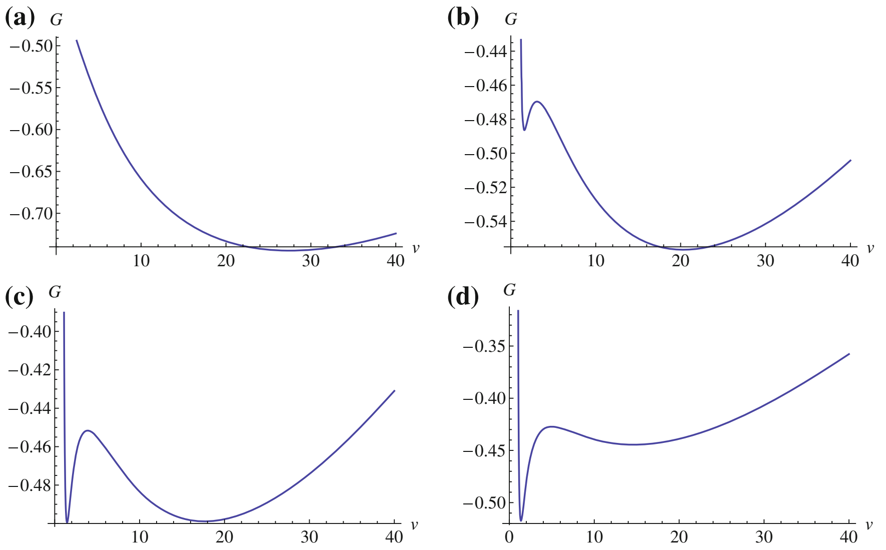


Fig. 3.5 Plots of the Gibbs free energy for van der Waals theory, with the parameters $a = 1$ and $b = 1$, at the low pressure $p = 0.01$. **a** At temperature $k_B T = 0.3$, with only one minimum at $v \approx 27$ (stable gas phase). **b** At $k_B T = 0.24$, with the absolute minimum at $v \approx 20$ (stable gas phase) and a local minimum at $v \approx 1.6$ (metastable liquid phase). **c** At $k_B T = 0.22$, with two equally deep minima at $v \approx 18$ and $v \approx 1.6$ (gas–liquid transition). **d** At $k_B T = 0.2$, with the absolute minimum at $v \approx 1.4$ (stable liquid phase) and a local minimum at $v \approx 15$ (metastable gas phase) (Interactive version at http://www.springer.com/cda/content/document/cda_downloaddocument/Selinger+Interactive+Figures.zip?SGWID=0-0-45-1509169-p177545420.)

gas case in Fig. 3.2, with just one minimum at a high volume. In Fig. 3.5b, when the temperature is reduced, the plot becomes more interesting. The absolute minimum is still at a high volume, but there is now also a local minimum at a much lower volume. As the temperature is reduced further, the low-volume minimum becomes deeper. At the temperature shown in Fig. 3.5c, the two minima are equally deep, and hence the system is right on the boundary between a high-volume state and a low-volume state. As the temperature continues to decrease, as shown in Fig. 3.5d, the low-volume minimum becomes even deeper than the high-volume minimum. (I encourage you to explore this behavior for yourself using the interactive version of the figure.)

In all of these plots, the deepest (absolute) minimum corresponds to the stable equilibrium phase of the system. The other (local) minimum corresponds to a metastable state; the concept of metastability will be discussed further in Chap. 7. Hence, we can see that the system is showing an abrupt transition from a gas state with a high volume per molecule at high temperature, to some other state with a low volume per molecule at low temperature. How can we interpret this behavior? What state is similar to a gas, but has a much lower volume per molecule (i.e., higher density) and forms at lower temperature? It is a liquid!

At this point, we can scan through pressure p and temperature T, and numerically find the volume per particle that minimizes the Gibbs free energy at each (p, T). If

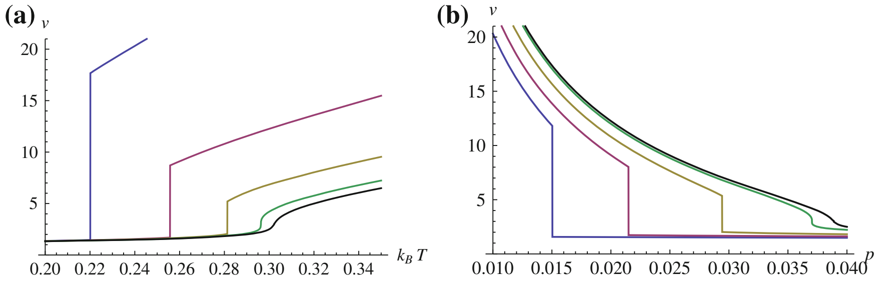


Fig. 3.6 Plots of the equilibrium volume per particle for van der Waals theory, with parameters $a = 1$ and $b = 1$. **a** Isobars at $p = 0.01$, $p = 0.02$, $p = 0.03$, $p = 1/27 \approx 0.037$, and $p = 0.04$ (from *left* to *right*). **b** Isotherms at $k_B T = 0.24$, $k_B T = 0.26$, $k_B T = 0.28$, $k_B T = 8/27 \approx 0.296$, and $k_B T = 0.30$ (from *left* to *right*)

the Gibbs free energy has two minima at some (p, T), we will just identify the deepest minimum. Figure 3.6a shows a series of plots as a function of temperature at constant pressure; these plots are called *isobars*. The first, blue line shows the behavior at low pressure (the same pressure shown in the Gibbs free energy plots of Fig. 3.5). In this case, there is an transition from the liquid phase at low temperature to the gas phase at high temperature. The volume per molecule of the gas phase is very sensitive to temperature, while the volume per molecule of the liquid phase is almost independent of temperature. The second and third lines show the behavior at higher pressures. We see that the transition temperature increases as the pressure increases (think of boiling water at lower or higher altitude). We also see that the volume discontinuity becomes smaller as the pressure increases, i.e., the difference between liquid and gas becomes smaller. Eventually, at the special pressure shown in the fourth, green line, the discontinuity between liquid and gas goes to zero. Instead of a discontinuity, the curve just shows a single point where the slope $\partial v/\partial T \to \infty$. This special point is called the *critical point;* it is quite analogous to the critical point for the Ising model, and it will be discussed in more detail in the next section. For even higher pressure, the fifth line has neither a discontinuity nor a point of infinite slope; it just shows a smooth change in the volume as a function of temperature. At this high pressure, the system cannot be identified as either liquid or gas; it is just called a *supercritical fluid,* meaning beyond the critical point.

Figure 3.6b shows a corresponding series of plots as a function of pressure at constant temperature; these plots are called *isotherms*. The first, blue line shows the behavior at low temperature. As the pressure increases, the system abruptly condenses from gas to liquid. The volume per molecule of the gas phase is very sensitive to pressure, while the volume per molecule of the liquid phase is almost independent of pressure. The second and third lines show the behavior at higher temperatures: As the temperature increases, the pressure required to condense the gas into a liquid increases, and the difference between gas and liquid becomes smaller. At the special pressure shown in the fourth, green line, the discontinuity between liquid and gas goes to zero, and we again see the critical point. Now it appears as a point with

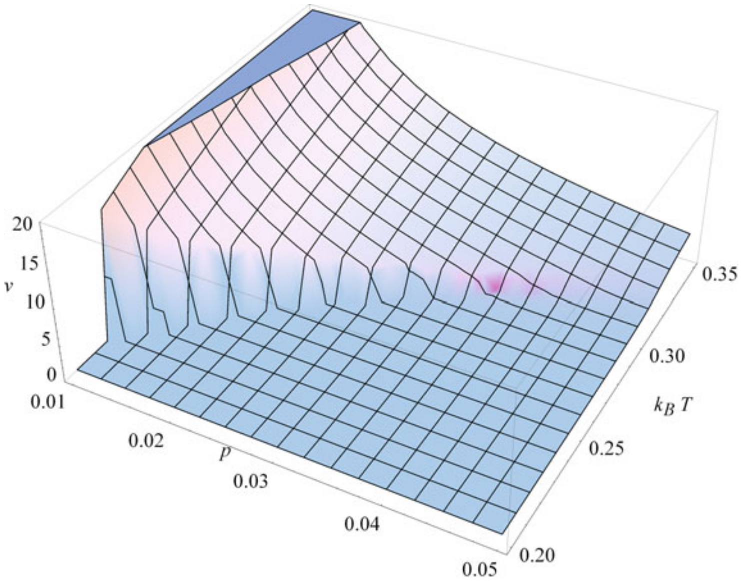


Fig. 3.7 Volume per particle as a function of temperature and pressure, in a 3D plot, for van der Waals theory with parameters $a = 1$ and $b = 1$ (Interactive version at http://www.springer.com/cda/content/document/cda_downloaddocument/Selinger+Interactive+Figures.zip?SGWID=0-0-45-1509169-p177545420.)

slope $\partial v/\partial T \to -\infty$. For even higher temperature, the fifth line has neither a discontinuity nor a point of infinite slope, but just a smooth change in volume as a function of pressure. Again we have the supercritical fluid.

Figure 3.7 shows a 3D plot of the volume per particle as a function of both temperature and pressure, which combines the information in both isobars and isotherms. It is quite analogous to the 3D plot of the Ising order parameter as a function of temperature and magnetic field in Fig. 2.8. Again, it looks like a partially torn sheet of paper. The “torn” part on the left represents the discontinuous volume change between liquid and gas, for temperature and pressure below the critical point. The “non-torn” part on the right represents the smooth response to temperature and pressure in the supercritical fluid regime.

If we project this 3D plot down into the (p, T) plane, we obtain the phase diagram shown in Fig. 3.8. It is analogous to the Ising phase diagram in Fig. 2.9. In this diagram, we see the gas phase at high temperature and low pressure, and the liquid phase at low temperature and high pressure. The sharp boundary between gas and liquid is a *first-order phase transition*. This boundary terminates at the critical point, where there is a *second-order phase transition*. This second-order transition does not have a discontinuity in the volume, but it has singularities in the derivatives $\partial v/\partial T$ and $\partial v/\partial p$. Beyond the critical point is the supercritical fluid regime, where gas and liquid cannot be distinguished. One can go from the gas phase to the liquid phase in two possible ways—either by crossing the transition line or by going around the critical point.

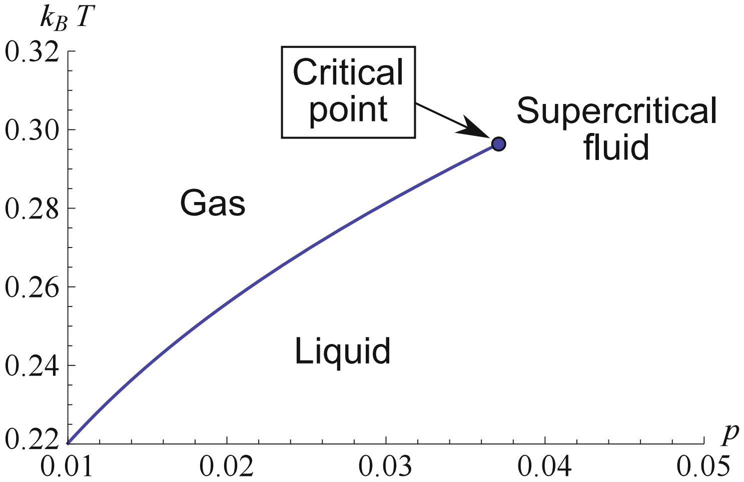


Fig. 3.8 Phase diagram as a function of pressure and temperature, for van der Waals theory with parameters $a = 1$ and $b = 1$

3.6 Behavior Near Critical Point

We have seen that the critical point is a special point in the van der Waals phase diagram, where the first-order liquid–gas transition comes to an end and the phases merge into a single supercritical fluid. In that respect, it is analogous to the critical point in the Ising phase diagram, where the first-order transition between the spin-up and spin-down phases comes to an end and they merge into a single paramagnetic phase. It is worthwhile to examine the behavior near the critical point in more detail. We will be able to make some analytic predictions here, without the need for numerical calculations.

For this analysis, let us return to plots of the Gibbs free energy as a function of volume per molecule. In Fig. 3.5 discussed above, we showed plots at low pressure, far from the critical point. Now, suppose we increase the pressure and temperature, and continue to track the gas–liquid transition. As pressure and temperature approach the critical point, as shown in Fig. 3.9a, b, the liquid and gas minima of $G(v)$ gradually come together. Exactly at the critical point, shown in Fig. 3.9c, the distinction between gas and liquid completely goes away, and the plot of the Gibbs free energy just has a single very flat minimum. Beyond the critical point, as in Fig. 3.9d, the curvature at the minimum becomes greater, and it looks more like a parabola. This series of plots is similar to the series of free energy plots for the Ising model shown in Fig. 2.4, except that the Ising plots have a symmetry between positive and negative M while the van der Waals plots do not have such a symmetry.

To calculate the pressure, temperature, and volume at the critical point, we note that the plot of $G(v)$ has a *quartic* minimum at that point. Because of the quartic minimum, we must have

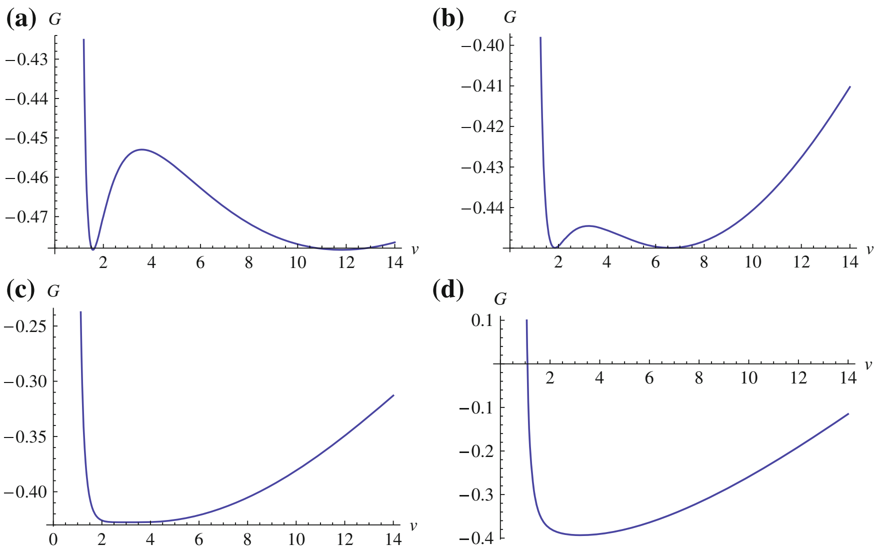


Fig. 3.9 Plots of the Gibbs free energy for van der Waals theory, with the parameters $a = 1$ and $b = 1$, near the critical point. **a** At $k_B T = 0.24$ and $p = 0.015$. **b** At $k_B T = 0.27$ and $p = 0.025$. **c** At the critical temperature $k_B T_C = 8/27 \approx 0.296$ and critical pressure $p_C = 1/27 \approx 0.037$. **d** In the supercritical regime at $k_B T = 0.35$ and $p = 0.061$ (Interactive version at http://www.springer.com/cda/content/document/cda_downloaddocument/Selinger+Interactive+Figures.zip?SGWID=0-0-45-1509169-p177545420.)

$$\frac{\partial G}{\partial v} = 0, \quad \frac{\partial^2 G}{\partial v^2} = 0, \quad \frac{\partial^3 G}{\partial v^3} = 0, \tag{3.46}$$

at the critical point. Hence, we have a set of three simultaneous equations in three unknowns: the critical pressure p_C, the critical temperature T_C, and the critical volume per particle v_C. The solution is

$$p_C = \frac{a}{27b^2}, \quad k_B T_C = \frac{8a}{27b}, \quad v_C = 3b. \tag{3.47}$$

These equations are consistent with the numerical calculation of the critical point in Fig. 3.9c. Note especially the prediction for v_C: We know that the minimum volume per particle at infinite pressure is b, so the critical volume per particle is three times that value.

Now suppose we are at a pressure and temperature slightly less than the critical pressure, $p = p_C + \delta p$ and $T = T_C + \delta T$, with δp and δT both negative. In this case, $G(v)$ has two minima at v slightly above and below v_C, as shown in Fig. 3.9b. Hence, we write $v = v_C + \delta v$, and expand the $G(v)$ as a power series to fourth order in δv,

$$\frac{G(v)}{N} = \frac{G_C}{N} + \left(\delta p - \frac{k_B \delta T}{2b}\right)\delta v + \frac{k_B \delta T}{8b^2}(\delta v)^2 \\ - \frac{k_B \delta T}{24b^3}(\delta v)^3 + \frac{8a + 243bk_B\delta T}{15552b^5}(\delta v)^4 + \cdots . \tag{3.48}$$

Let us assume that the minima have δv of order $|\delta T|^{1/2}$ and $|\delta p|^{1/2}$; we will verify that assumption self-consistently in a moment. In this case, we can neglect the cubic term and the second quartic term in comparison with the other terms, and the series simplifies to

$$\frac{G(v)}{N} \approx \frac{G_C}{N} + \left(\delta p - \frac{k_B \delta T}{2b}\right)\delta v + \frac{k_B \delta T}{8b^2}(\delta v)^2 + \frac{8a}{15552b^5}(\delta v)^4 + \cdots . \tag{3.49}$$

The liquid–gas transition occurs when the two minima with positive and negative δv have equal Gibbs free energies. This occurs when the coefficient of the linear term in δv is zero:

$$\delta p = \frac{k_B \delta T}{2b}. \tag{3.50}$$

At this transition, the gas and liquid phases have the volumes per particle given by the minima of $G(v)$, which occur at

$$\delta v = \pm \left|\frac{243b^3 k_B \delta T}{2a}\right|^{1/2}, \tag{3.51}$$

so our assumption about the scaling of δT is self-consistent.

The conclusion of this analysis can be written as a simple power law: The difference between the gas and liquid volumes scales as

$$v_L - v_G \propto (T_C - T)^{\beta}, \tag{3.52}$$

where β is a *critical exponent*. Here, we find $\beta = \frac{1}{2}$ in mean-field theory for the gas–liquid transition. This result is the same exponent as mean-field theory for the Ising model! In Chap. 4, we will discuss the reason for this similarity.

Problem: The *isothermal compressibility* of a material is defined as

$$\kappa_T = -\frac{1}{V}\left(\frac{\partial V}{\partial p}\right)_T, \tag{3.53}$$

where the derivative is taken at constant temperature. In the supercritical fluid, with T slightly above T_C, how does the isothermal compressibility depend on $(T - T_C)$?

Solution: In the supercritical fluid, close to the critical point, δv is very small. Hence, we can approximate the Gibbs free energy by the series to second order in δv:

$$\frac{G}{N} \approx \frac{G_C}{N} + \left(\delta p - \frac{k_B \delta T}{2b}\right)\delta v + \frac{k_B \delta T}{8b^2}(\delta v)^2. \tag{3.54}$$

Minimizing this expression over δv gives

$$0 = \frac{\partial}{\partial(\delta v)}\left[\frac{G}{N}\right] \approx \delta p - \frac{k_B \delta T}{2b} + \frac{k_B \delta T}{4b^2}\delta v, \tag{3.55}$$

and hence

$$\delta v = 2b\left(1 - \frac{2b\delta p}{k_B \delta T}\right). \tag{3.56}$$

The response to a change in pressure is

$$\frac{\partial(\delta v)}{\partial p} = -\frac{4b^2}{k_B \delta T}. \tag{3.57}$$

Thus, the isothermal compressibility is

$$\kappa_T = -\frac{1}{V}\left(\frac{\partial V}{\partial p}\right)_T = -\frac{1}{v_C}\frac{\partial(\delta v)}{\partial p} = \frac{1}{3b}\frac{4b^2}{k_B \delta T} = \frac{4b}{3k_B(T - T_C)}. \tag{3.58}$$

This dependence on $(T - T_C)$ can be written as the scaling relation

$$\kappa_T \propto (T - T_C)^{-\gamma}, \tag{3.59}$$

with the critical exponent $\gamma = 1$. This is the same exponent as mean-field theory for the susceptibility of the Ising model in the paramagnetic phase!

3.7 Gas–Liquid Transition at Fixed Volume

As a final point about the gas–liquid transition, we must consider one more issue: Experiments are not always done at fixed pressure. Instead, they are sometimes done at *fixed volume*.

Suppose we put N molecules into a container with volume V. We imagine that this container has very strong walls, so it cannot possibly change its volume. We control the temperature T of the system, perhaps by putting the whole container into an oven. In this case, we do not control the pressure p; instead, the system makes its own pressure inside the container. We can now ask two questions: First, what is the phase inside the container: gas or liquid? Second, what is the pressure inside the container?

Let us begin with a specific example of van der Waals theory with parameters $a = 1$ and $b = 1$. Suppose the temperature is $k_B T = 0.24$, which is the same temperature as the first, blue isotherm in Fig. 3.6b. We can consider three cases:

(a) Suppose we choose N and V such that $V/N = 13$. From the isotherm at $k_B T = 0.24$, we see that the gas phase at $p = 0.014$ has the right volume $v = 13$. Hence, we can conclude that the container is filled with gas phase at $p = 0.014$. In general, for any $V/N > 12$, we can find *some pressure* such that the gas phase at that pressure has the volume we are looking for. Thus, for any $V/N > 12$, we can conclude that the container is filled with gas phase at the corresponding pressure.

(b) Suppose we choose N and V such that $V/N = 1.5$. From the isotherm at $k_B T = 0.24$, we see that the liquid phase at $p = 0.035$ has the right volume $v = 1.5$. Hence, we can conclude that the container is filled with liquid phase at $p = 0.035$. In general, for any V/N between 1.0 (the minimum possible volume of the system) and 1.6 (the maximum volume of the liquid phase at $k_B T = 0.24$), we can find *some pressure* such that the liquid phase at that pressure has the volume we are looking for. Thus, for any V/N between 1.0 and 1.6, we can conclude that the container is filled with liquid phase at the corresponding pressure. (Of course, if we want V/N to be close to 1.0, we might need to go to an enormous pressure, but we will not worry about that practical consideration here.)

(c) Suppose we choose N and V such that $V/N = 5$. From the isotherm at $k_B T = 0.24$, we see that the liquid phase has a maximum volume of 1.6, and the gas phase has a minimum volume of 12. Neither phase has a volume per molecule of 5 at any pressure; this volume per molecule is right in the middle of the discontinuity between liquid and gas. So what happens inside the container with $V/N = 5$?

To answer this question, let us go back to the Gibbs free energy plot at $k_B T = 0.24$, right at the transition pressure of $p = 0.015$; this plot is shown in Fig. 3.9a. On this plot, we see that the selected volume per molecule, $V/N = 5$, is between the liquid minimum at $v = 1.6$ and the gas minimum of $v = 12$. There exists a state with $v = 5$, but this state has a higher Gibbs free energy than $v = 1.6$ or $v = 12$. Hence, the system can reduce its Gibbs free energy by breaking up into one region of $v = 1.6$ and another region of $v = 12$. This is exactly what happens in the experiment! Inside the container, there is not any single uniform phase; instead, there is a *coexistence* of some liquid at $v = 1.6$ and some gas at $v = 12$.

In general, the behavior for fixed volume per molecule and fixed temperature is shown in the phase diagram of Fig. 3.10. At high volume, the system is entirely in the gas phase. This single gas phase exists down to a minimum volume per molecule at that temperature (which is the same minimum v for the gas phase shown in the isotherm). At low volume, the system is entirely in the liquid phase. This single liquid phase exists up to a maximum volume per molecule at that temperature (which is the same maximum v for the liquid phase shown in the isotherm). In between the minimum v for the gas and the maximum v for the liquid, the system shows *two-phase coexistence* of gas and liquid, sometimes called a *biphasic* region. In the phase diagram, the two-phase coexistence region is indicated by horizontal lines, which remind us that the two coexisting phases have the same temperature but different v. As the temperature increases toward the critical point, the distinction between

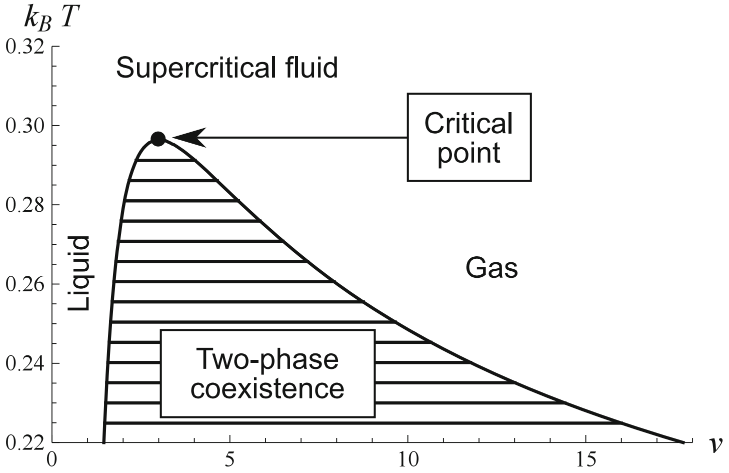


Fig. 3.10 Phase diagram as a function of volume per molecule and temperature, for van der Waals theory with parameters $a = 1$ and $b = 1$

gas and liquid becomes smaller, and hence the width of the two-phase coexistence region becomes smaller. At the critical point, the gas and liquid become identical and the two-phase coexistence vanishes. Above the critical temperature, the system is entirely the supercritical fluid, which cannot be identified as either liquid or gas.

If the system is in the two-phase coexistence region, so that it consists of some gas and some liquid, then we might want to know how much gas and how much liquid is in the container. Suppose the gas phase contains N_G molecules in a volume V_G, while the liquid phase contains N_L molecules in a volume V_L. We then have four equations in these four unknowns:

$$\begin{aligned} &N_G + N_L = N, \quad V_G + V_L = V, \\ &\frac{V_G}{N_G} = v_G(T), \quad \frac{V_L}{N_L} = v_L(T). \end{aligned} \tag{3.60}$$

The first two equations express the fact that the total number of molecules in each phase must add up to the total number of molecules in the container, and the total volume in each phase must add up to the total volume in the container. The last two equations express the fact that the volume per molecule in each phase must be the appropriate functions of temperature—the minimum volume per molecule of the gas phase and the maximum volume per molecule of the liquid phase—as shown on the phase diagram. In order for the system to be in the two-phase coexistence region, we must have $v_L(T) < (V/N) < v_G(T)$.

The solution of these four simultaneous equations is

$$\begin{aligned} N_G &= \frac{V - v_L(T)N}{v_G(T) - v_L(T)}, \quad V_G = v_G(T)\left(\frac{V - v_L(T)N}{v_G(T) - v_L(T)}\right), \\ N_L &= \frac{v_G(T)N - V}{v_G(T) - v_L(T)}, \quad V_L = v_L(T)\left(\frac{v_G(T)N - V}{v_G(T) - v_L(T)}\right). \end{aligned} \tag{3.61}$$

This solution is called the *lever rule*, because it is analogous to the balance of torques on a lever (when a large mass is near the fulcrum, and a small mass is far from the fulcrum). We should notice three features of the solution:

First, when the system is near the gas side of the coexistence region, then $V/N \to v_G(T)$, and hence $N_G \to N$, $N_L \to 0$, $V_G \to V$, and $V_L \to 0$, so the system is almost all gas. Likewise, when the system is near the liquid side of the coexistence region, then $V/N \to v_L(T)$, and hence $N_G \to 0$, $N_L \to N$, $V_G \to 0$, and $V_L \to V$, so the system is almost all liquid. Between these limits, the composition goes smoothly between mostly gas and mostly liquid.

Second, whenever we say that a system is some percentage gas or some percentage liquid, we must say whether that is by number or by volume. The percentage gas by number is

$$\frac{N_G}{N} = \frac{(V/N) - v_L(T)}{v_G(T) - v_L(T)}, \tag{3.62}$$

while the percentage gas by volume is

$$\frac{V_G}{V} = \left(\frac{v_G(T)}{V/N}\right)\left(\frac{(V/N) - v_L(T)}{v_G(T) - v_L(T)}\right), \tag{3.63}$$

and likewise for the percentage liquid. The percentages by number and by volume may be quite different. Indeed, it is quite common for most of the molecules in a system to be in the liquid phase, while most of the volume is in the gas phase, because $v_G(T)$ is much greater than $v_L(T)$. This situation should be familiar to chemistry students, who must specify the percentage of chemicals by mole, by volume, or by weight; those percentages are all different.

Third, suppose we vary the temperature of a system. The number of molecules N and volume V are fixed; they do not depend on temperature. However, the minimum gas volume $v_G(T)$ and maximum liquid volume $v_L(T)$ do depend on temperature, as shown in the phase diagram of Fig. 3.10. Hence, the percentages of the system that are gas and liquid must vary with temperature.

Figure 3.11 shows two examples of what can happen. In the top row, we have a system with V/N greater than the critical volume per molecule v_c. At low temperature, it has a coexistence between gas and liquid, with percentages given by the lever rule. As the temperature increases, $v_G(T)$ decreases toward the actual V/N for the system, and hence the percentage of gas increases. At some temperature, $v_G(T)$ becomes equal to V/N, and hence the system becomes entirely gas. It remains entirely gas for higher temperature. This is exactly the process of *boiling* a liquid at fixed volume.

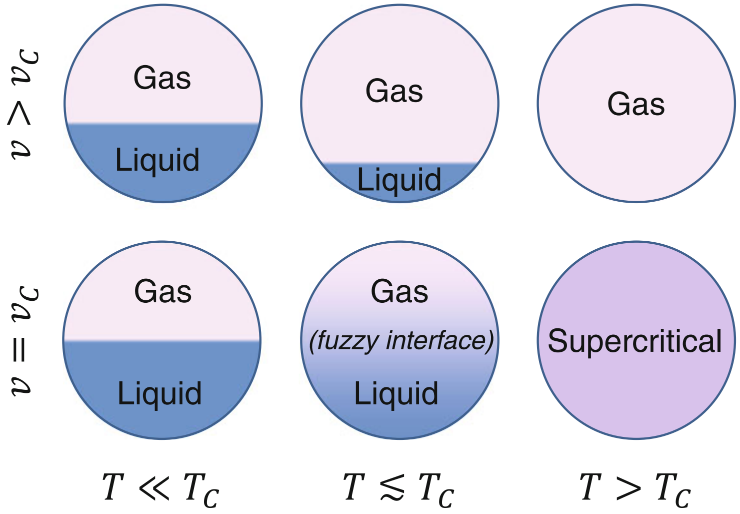


Fig. 3.11 Examples of the liquid–gas transition at fixed volume, as the temperature is increased. The *top* row shows a high volume per molecule (low density), while the *bottom* row shows the critical volume per molecule (critical density)

By comparison, in the bottom row, we have a system with V/N equal to v_C. At low temperature, it also has a coexistence between gas and liquid, with percentages given by the lever rule. As the temperature increases, this V/N remains between $v_G(T)$ and $v_L(T)$. The percentages of liquid and gas may change somewhat, but the system does not become entirely gas or liquid; it continues to have some gas and some liquid. As the temperature approaches T_C, the distinction between gas and liquid becomes smaller, and the interface between them gradually becomes fuzzy and diffuse. (The fuzzy interface will be discussed further in Chap. 6.) At the critical point, the interface completely blurs out; gas and liquid can no longer be distinguished. For all temperatures beyond T_C, the system is entirely the supercritical fluid.

Problem: Suppose we have a shipping container of size 12.0 m × 2.35 m × 2.38 m at room temperature. We pump out all the air, and put 1,000 kg of water into the container. How much water will be in the liquid phase, and how much in the gas phase?

Solution: The total volume of the container is $V = V_L + V_G = 67\text{ m}^3$, with V_L in the liquid phase and V_G in the gas phase. The total mass of water is $M = M_L + M_G = 1,000\text{ kg}$.

The density of liquid water is well known to be $M_L/V_L = 1\text{ g/cm}^3 = 1000\text{ kg/m}^3$. The maximum density of water vapor is not as well known, but it can be found in tables of the thermodynamic properties of water, which engineers call *steam tables*.[3] At room temperature of $T = 22.5\,^\circ\text{C}$, these tables give $M_G/V_G = 0.02\text{ kg/m}^3$.

[3] For example, see http://www.efunda.com/materials/water/steamtable_sat.cfm.

Solving these equations simultaneously gives $V_L = 0.9987\,\mathrm{m}^3$, $V_G = 66.0013\,\mathrm{m}^3$, $M_L = 998.7$ kg, and $M_G = 1.3$ kg. Hence, almost all of the volume is in the gas phase, but almost all of the mass is in the liquid phase. This situation is very common, because the density of liquid water is much greater than the density of water vapor.

In this solution, we have worked in terms of volume and mass, rather than volume and number of water molecules. We could have converted from mass to number of molecules, but this conversion is not necessary in a system of pure water. Mass is equivalent to number of molecules, because each water molecule has a constant mass.

Further Reading

Van der Waals theory for gases and liquids is discussed in many textbooks on statistical mechanics and physical chemistry. Some examples are:

1. D.L. Goodstein, *States of Matter* (Prentice-Hall, 1975)
2. T.L. Hill, *An Introduction to Statistical Thermodynamics* (Addison-Wesley, 1960)
3. S.-K. Ma, *Statistical Mechanics* (World Scientific, 1985)

Chapter 4
Landau Theory

Abstract As an alternative to the microscopic theories presented previously, this chapter presents Landau theory, which is a very general approach to understanding phase transitions. Landau theory ignores the microscopic structure of matter; instead, it is based on considerations of symmetry and the smoothness of functions. Remarkably, it gives predictions for phase transitions that are similar to predictions from microscopic theories.

So far we have discussed theories for two types of phase transitions: the Ising model for ferromagnetism and van der Waals theory for the gas-liquid transition. These theories are quite different—one is based on spins on a lattice, while the other is based on molecules with positions and momenta. However, we have seen that the predictions of these theories have a remarkable similarity. They both predict a first-order transition with a discontinuity in some order parameter, the magnetization for the Ising model and the volume per molecule for van der Waals theory. In both cases, the first-order transition terminates in a critical point, where the magnitude of the discontinuity goes to zero. Just below the critical point, the order parameter scales as $(T_C - T)^\beta$, with the critical exponent $\beta = \frac{1}{2}$.

At this point, you might be wondering: Is there any fundamental reason why the predictions of these theories are so similar? Can there be any unified description for both of these transitions, and perhaps also for other phase transitions?

Of course, the answer to this question is yes. There is a general approach that provides information about these two phase transitions, as well as many others. It does not provide all the information that comes from more specific theories, like the Ising model and van der Waals theory, but it at least provides information about the structure of phase diagrams and about critical exponents. This approach is called Landau theory, after the great Soviet theoretical physicist Lev Landau.

Personally, I like to think of Landau theory as "how to get something for nothing." Everyone in the world wants to get something for nothing—whenever I see a free sample in the supermarket, I want to pick it up. Landau theory is the physics version of that impulse. Here, we do not assume anything about the structure of the world; we totally ignore the fact that a magnet is made of microscopic spins, or a fluid is made of atoms and molecules. Instead, we just think about the symmetry of order

J.V. Selinger, *Introduction to the Theory of Soft Matter*,
Soft and Biological Matter, DOI 10.1007/978-3-319-21054-4_4

parameters, and see what we can infer from that. You might not expect that we can infer anything from such a simple starting point. It is truly amazing to me that this approach works so well.

4.1 Ising Ferromagnetism

Let us begin with the example of an Ising magnet, i.e., a magnet that can only have magnetic order along one axis (call it $\hat{z}$). For the moment, let us consider the case with no applied field. We want to develop a theory for the transition from the disordered paramagnetic state to the ordered ferromagnetic state. Suppose we do not know anything about the microscopic structure of the magnet; we only know that it has an order parameter M, which is zero in the paramagnetic state and becomes nonzero in the ferromagnetic state. Hence, we must make a model for the free energy as a function of M. Now we make a crucial assumption: We assume that the free energy is a smooth[1] function of M around $M = 0$. Why is that assumption reasonable? Well, the basic concept of the theory is that we begin with assumptions about smoothness, and we end by making predictions about phase transitions that are *not* smooth; they are discontinuous or singular changes. The *input* into the theory must be smooth, and then it is remarkable that the *output* of the theory is not smooth. If the input were not a smooth function, then the input could be anything, and hence the output could be anything. It is only this assumption of smooth input that puts a physical constraint on the output.

Because we assume that F is a smooth function of M, we can expand F as a *power series* about $M = 0$,

$$F = F_0 + F'(0)M + \frac{1}{2!}F''(0)M^2 + \frac{1}{3!}F'''(0)M^3 + \frac{1}{4!}F''''(0)M^4 + \cdots . \quad (4.1)$$

We assume that higher-order terms are negligible for small M, sufficiently close to the paramagnetic–ferromagnetic transition. Now we consider the symmetry of the system. We know that the system has a symmetry under reflection in the xy-plane, which changes $\hat{z} \to -\hat{z}$. For this reason, the free energy must be unchanged under this reflection, so that $F(M) = F(-M)$. In other words, F must be an *even* function of M. For this reason, the odd terms in the power series must vanish, $F'(0) = F'''(0) = 0$. The remaining terms are

$$F = F_0 + \frac{1}{2!}F''(0)M^2 + \frac{1}{4!}F''''(0)M^4. \quad (4.2)$$

At this point, just for future convenience, we make two small changes in notation. First, because the free energy is extensive, it should be proportional to the volume.

[1] If you know complex analysis, the assumption is that F is an *analytic* function of M. If you do not know complex analysis, you can just use the word smooth.

Hence, we will write the series for the free energy *density*, i.e., the free energy per volume. Second, we relabel the series coefficients in the more compact form:

$$f = \frac{F}{V} = f_0 + \frac{1}{2}aM^2 + \frac{1}{4}bM^4. \tag{4.3}$$

Now we must consider how the free energy depends on temperature. In principle, all of the coefficients f_0, a, and b must be functions of temperature, so we can write the power series as

$$f = f_0(T) + \frac{1}{2}a(T)M^2 + \frac{1}{4}b(T)M^4. \tag{4.4}$$

Of course, we do not know how they depend on temperature. We will only assume that the dependence on temperature is *smooth*, for the same reason discussed above. Hence, all of them can be expanded as power series about any arbitrary temperature T_0,

$$a(T) = a_0 + a'(T - T_0) + \cdots, \quad b(T) = b_0 + b'(T - T_0) + \cdots. \tag{4.5}$$

The same is also true for $f_0(T)$, but it will not really matter. In a moment, we will see that the interesting behavior happens near the temperature where $a(T)$ passes through 0. Hence, we will choose the arbitrary temperature T_0 such that $a(T_0) = a_0 = 0$. For temperatures T near T_0, we can then assume

$$a(T) = a'(T - T_0). \tag{4.6}$$

Likewise, for temperatures T near T_0, we can neglect the linear term in $b(T)$ and approximate it by the constant term

$$b(T) = b_0. \tag{4.7}$$

Hence, our series for the free energy density becomes

$$f = f_0 + \frac{1}{2}a'(T - T_0)M^2 + \frac{1}{4}b_0M^4. \tag{4.8}$$

Figure 4.1 shows plots of this free energy for temperatures $T > T_0$, $T = T_0$, and $T < T_0$. We can see that it has the same general form as the free energy in Fig. 2.4, which we determined from microscopic mean-field theory. As the temperature decreases from $T > T_0$ to $T < T_0$, the minimum at $M = 0$ gradually flattens out and then splits into two minima at positive and negative M. This splitting of the minimum is precisely the spontaneous symmetry-breaking transition from the paramagnetic phase to the ferromagnetic phase at the Ising critical point. Hence, we can identify the parameter T_0 in the series expansion with the Ising critical temperature T_C, and we will use that symbol.

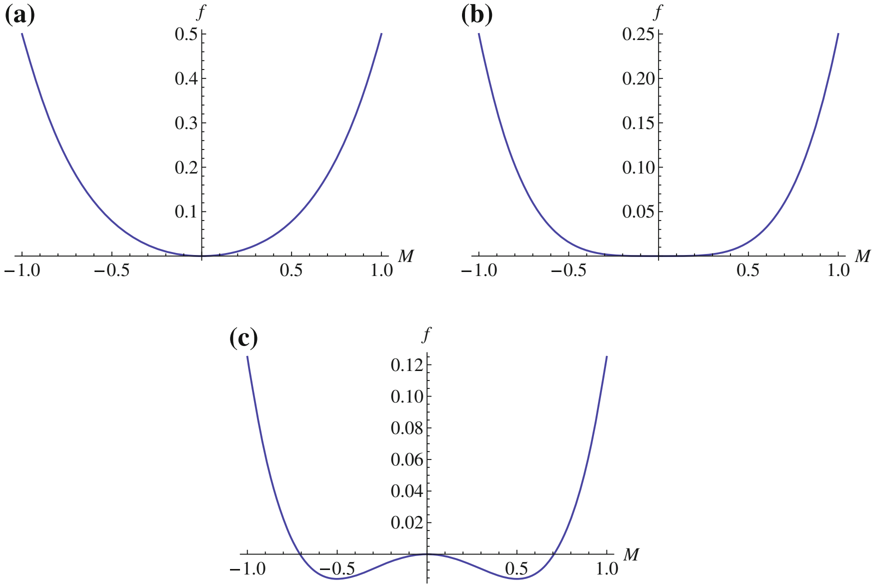


Fig. 4.1 Plots of the free energy of Eq. (4.8) from Landau theory for an Ising ferromagnet. **a** $T > T_0$. **b** $T = T_0$. **c** $T < T_0$ (Interactive version at http://www.springer.com/cda/content/document/cda_downloaddocument/Selinger+Interactive+Figures.zip?SGWID=0-0-45-1509169-p177545420.)

To find the minimum of this free energy, we set the first derivative equal to zero:

$$\frac{\partial f}{\partial M} = a'(T - T_C)M + b_0 M^3 = 0. \tag{4.9}$$

For $T \geq T_C$, the only real solution is $M = 0$, corresponding to the disordered, paramagnetic phase. For $T < T_C$, there are three real solutions:

$$M = 0, \quad M = \pm\left(\frac{a'(T_C - T)}{b_0}\right)^{1/2}. \tag{4.10}$$

Of these three solutions, $M = 0$ is the local *maximum* shown in Fig. 4.1c, while the other two solutions are the positive and negative minima. The behavior as a function of temperature is shown in Fig. 4.2. We can see that it follows the scaling relation

$$M \propto (T_C - T)^{\beta}, \tag{4.11}$$

where the critical exponent is $\beta = \frac{1}{2}$. Hence, Landau theory even gives the same critical exponent as we found in Eq. (2.38), based on microscopic mean-field theory for the Ising model. It is remarkable that Landau theory can provide this much information based on nothing but symmetry arguments!

This Landau theory can be generalized by adding the effects of a magnetic field h. Any applied field breaks the symmetry between M and $-M$, and hence the free

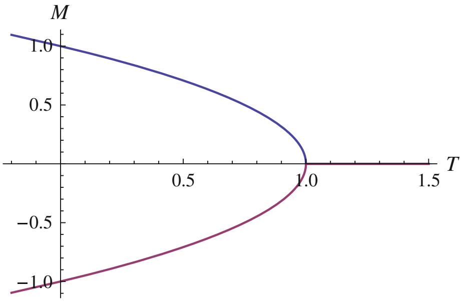


Fig. 4.2 Prediction of Landau theory for the magnetic order parameter as a function of temperature

energy series can have odd as well as even terms in M. Let us suppose that h is small, so that it is just a perturbation on the symmetric system. For small h and small M, the lowest-order coupling between h and M is just a term of $-hM$, with some arbitrary coefficient μ. There could certainly be additional couplings like $-h^3M$ and $-hM^3$, but they are much smaller because they involve higher powers of h and M. Hence, we can write the leading terms in the free energy series as

$$f = f_0 - \mu hM + \frac{1}{2}a'(T - T_C)M^2 + \frac{1}{4}b_0M^4. \tag{4.12}$$

This expression allows us to predict the effects of a small symmetry-breaking field on the order parameter, again based on nothing but symmetry arguments.

Problem: Using Landau theory for $T > T_C$, show that the susceptibility defined by Eq. (2.42) diverges as the system approaches the critical point, with the scaling relation $\chi \propto (T - T_C)^{-\gamma}$, where $\gamma = 1$. This is the same critical exponent as derived from the microscopic mean-field theory of Chap. 2.

Solution: For $T > T_C$, if the applied field is small, the induced order parameter must also be small. Hence, we can neglect the $\frac{1}{4}b_0M^4$ term in comparison with the $\frac{1}{2}a'(T - T_C)M^2$ term in the free energy, and approximate the free energy by

$$f = f_0 - \mu hM + \frac{1}{2}a'(T - T_C)M^2. \tag{4.13}$$

To minimize this free energy over M, we set $\partial f/\partial M = 0$, and hence obtain

$$M = \frac{\mu h}{a'(T - T_C)}. \tag{4.14}$$

The susceptibility is therefore

$$\chi = \frac{\partial M}{\partial h} = \frac{\mu}{a'(T - T_C)}, \tag{4.15}$$

which follows the scaling relation with critical exponent $\gamma = 1$.

Problem: Using Landau theory for $T = T_C$, show that the response to a small applied field follows the scaling relation $M \propto h^{1/\delta}$, where $\delta = 3$. Again, this is the same critical exponent as in microscopic mean-field theory.

Solution: At $T = T_C$, the free energy reduces to

$$f = f_0 - \mu h M + \frac{1}{4} b_0 M^4. \tag{4.16}$$

We minimize this expression by setting $\partial f / \partial M = 0$, and obtain

$$M = \left(\frac{\mu h}{b_0} \right)^{1/3}, \tag{4.17}$$

which follows the scaling relation with exponent $\delta = 3$.

Landau theory can also be generalized to consider a *nonuniform* order parameter $M(\boldsymbol{r})$, which depends on position. In this case, the free energy density $f(M(\boldsymbol{r}))$ also depends on position. Because of this dependence, the total free energy is not just the volume times the free energy density. Instead, it must be the *integral* of the free energy density over position:

$$F = \int d^3 r f(M(\boldsymbol{r})). \tag{4.18}$$

For a nonuniform system, the free energy density includes the series of Eq. (4.12), and it may also include an extra free energy cost associated with the variation of $M(\boldsymbol{r})$. What is this extra free energy cost? Well, we have to assume that $M(\boldsymbol{r})$ is a slowly varying function of position, which only changes over long length scales. For this reason, the first derivatives are small, and the higher derivatives are even smaller. Hence, the main contribution to the extra free energy cost must come from the gradient ∇M. How can the extra free energy cost depend on ∇M? Well, ∇M is a vector but f is a scalar. We need to construct a scalar that depends on the vector in a smooth way. Once again, we make a power series, now a series in ∇M. The lowest-order, largest term in this power series must be proportional to $|\nabla M|^2$, with some arbitrary coefficient $\frac{1}{2} K$. Hence, we write the gradient free energy as

$$f_{\text{gradient}} = \frac{1}{2} K |\nabla M|^2. \tag{4.19}$$

When we combine the gradient free energy with our previous series, we obtain the free energy density

$$f = f_0 - \mu h M + \frac{1}{2} a'(T - T_C) M^2 + \frac{1}{4} b_0 M^4 + \frac{1}{2} K |\nabla M|^2, \tag{4.20}$$

and hence the total free energy

$$F = \int d^3r \left[f_0 - \mu h M + \frac{1}{2} a'(T - T_C) M^2 + \frac{1}{4} b_0 M^4 + \frac{1}{2} K |\nabla M|^2 \right]. \quad (4.21)$$

This expression for the free energy, including the gradient term, will be useful when we consider interfaces in Chap. 6.

Before going on, I must acknowledge that Landau theory has some limitations compared with the mean-field theory presented in Chap. 2. (I promised that Landau theory gives you *something* for nothing, but it cannot give you *everything* for nothing!) We should note two points:

First, the parameters a' and b_0 in Landau theory are arbitrary coefficients; we do not have any physical interpretation of what they mean. By contrast, the parameters in our microscopic mean-field theory have specific physical meaning: J is the strength of the interaction between neighboring spins, and q is the coordination number of the lattice. In some cases, the generality of Landau theory is a real advantage, because the theory can apply to many different specific problems. In other cases, the generality is a disadvantage, because we do not know how the predictions are related to the microscopic structure of a physical system.

Second, Landau theory does not know that M can only be between $+1$ and -1. Indeed, Fig. 4.2 shows that the Landau prediction for M goes past ± 1 as the temperature decreases. Of course, we know that values of M beyond ± 1 are unphysical. The reason for this problem is that Landau theory is based on an expansion of the free energy for *small* M. It just does not apply when the magnitude of M grows large. Mathematically, the Landau free energy is based on a power series in M of Eq. (4.8), which does not do anything special when M reaches ± 1. By comparison, the microscopic mean-field free energy of Eq. (2.31) has singularities at $M = \pm 1$, and is undefined beyond those limits. For this reason, microscopic mean-field theory correctly predicts that M must saturate at ± 1, as shown in Fig. 2.6.

This second point leads to an important conclusion: *If you're using Landau theory, and your calculation predicts an order parameter that's unphysically large, then you're outside the regime where Landau theory is valid, so you shouldn't be using it.* Even so, Landau theory is extremely useful in the regime where it is valid: *when the order parameter is small.*

4.2 Gas-Liquid Transition

In the previous section, we constructed Landau theory for the Ising ferromagnet. Now let us construct Landau theory for the gas-liquid transition, and see how similar it is.

In Landau theory for gases and liquids, we assume that the free energy is a smooth function of the volume per molecule v. We will therefore expand it as a power series. However, we must now ask: A power series about what point? Should we expand it as a power series about $v = 0$? No, this would not work. The state with $v = 0$ is a truly pathological state with zero volume per molecule; it is essentially a black hole. We cannot expect to get any reasonable predictions by expanding about this state.

Alternatively, should we change variables to the density $\rho = 1/v$, and expand in a power series about $\rho = 0$? The concept of expanding about $\rho = 0$ is not crazy. The state with $\rho = 0$ is the ideal gas, which we understand well. However, the state with $\rho = 0$ is not close to the gas-liquid transition, so an expansion about that state is not a very effective way to learn about this transition.

A more effective approach would be to expand about the critical point itself. We will assume that a critical point exists at some temperature T_C, pressure p_C, and volume per molecule v_C. In other words, we assume that we can tune three parameters T, p, and v to find a point where three equations are satisfied:

$$\frac{\partial F}{\partial v} = 0, \quad \frac{\partial^2 F}{\partial v^2} = 0, \quad \frac{\partial^3 F}{\partial v^3} = 0. \tag{4.22}$$

We can then write the free energy density as a power series in $\delta v = v - v_C$,

$$f = \frac{F}{V} = f_0 + a(p,T)\delta v + \frac{1}{2}b(p,T)(\delta v)^2 + \frac{1}{3}c(p,T)(\delta v)^3 + \frac{1}{4}d(p,T)(\delta v)^4, \tag{4.23}$$

where a, b, c, and d are arbitrary coefficients. All of these coefficients may vary with pressure and temperature in some smooth way. At $p = p_C$ and $T = T_C$, we must have $a = b = c = 0$, while $d > 0$. Nearby, at $p = p_C + \delta p$ and $T = T_C + \delta T$, we can write their lowest-order dependence on pressure and temperature as

$$\begin{aligned}
a(p,T) &= a_p \delta p + a_T \delta T,\\
b(p,T) &= b_p \delta p + b_T \delta T,\\
c(p,T) &= c_p \delta p + c_T \delta T,\\
d(p,T) &= d_0 + d_p \delta p + d_T \delta T \approx d_0.
\end{aligned} \tag{4.24}$$

Now we can investigate what value of v minimizes the free energy. At the critical point, the free energy is a quartic function of δv, and the only minimum is $\delta v = 0$. Slightly away from the critical point, the free energy mainly depends on volume through the linear term $a\delta v$. If the coefficient $a < 0$, the favored state has $\delta v > 0$; if $a > 0$, the favored state has $\delta v < 0$. Hence, we will consider the borderline given by $a = 0$, or

$$\delta p = -\frac{a_T}{a_P}\delta T, \tag{4.25}$$

as shown in the pressure–temperature phase diagram of Fig. 4.3.

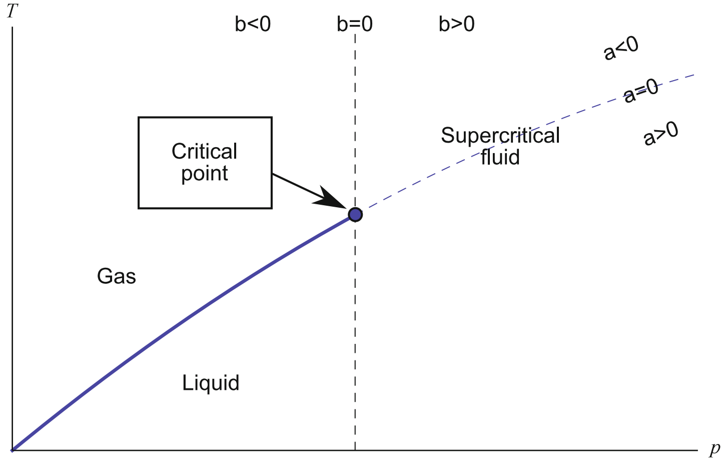


Fig. 4.3 Schematic phase diagram for the gas-liquid transition, derived from Landau theory

Along the $a = 0$ borderline, the free energy mainly depends on v through the quadratic term $\frac{1}{2}b(\delta v)^2$. Depending on the sign of b, two types of behavior are possible:

1. If $b < 0$, the leading contributions to the free energy along the $a = 0$ borderline are given by

$$f = \frac{F}{V} = f_0 - \frac{1}{2}|b|(\delta v)^2 + \frac{1}{4}d(\delta v)^4. \tag{4.26}$$

This form of the free energy is analogous to the Ising model at $T < T_C$, as discussed in the previous section. It has two minima at

$$\delta v = \pm \left|\frac{b}{d}\right|^{1/2} = \pm \left|\frac{b_p \delta p + b_T \delta T}{d_0}\right|^{1/2} = \pm \left|\frac{(b_T - b_p a_T/a_P)\delta T}{d_0}\right|^{1/2}. \tag{4.27}$$

Hence, there is a discontinuity between a gas phase with $v > v_C$ and a liquid phase with $v < v_C$. The magnitude of the discontinuity scales as $|T - T_C|^\beta$, with the critical exponent $\beta = \frac{1}{2}$.

2. If $b > 0$, the leading contributions to the free energy along the $a = 0$ borderline are

$$f = \frac{F}{V} = f_0 + \frac{1}{2}|b|(\delta v)^2 + \frac{1}{4}d(\delta v)^4. \tag{4.28}$$

This form of the free energy is analogous to the Ising model at $T > T_C$. It has a single minimum, which corresponds to the supercritical fluid state. With a little more work off of the $a = 0$ borderline, we can extract the critical exponent $\gamma = 1$, which characterizes the response to a change in pressure.

The conclusion of this argument is that we can reproduce some important features of van der Waals theory—including the structure of the phase diagram and the mean-field critical exponents—without actually using any of the microscopic assumptions of van der Waals theory. All we need is the general assumption that the free energy is a smooth function of δv, with coefficients that are smooth functions of δp and δT, near the critical point.

Of course, Landau theory does not give us all the information of van der Waals theory. In particular, it does not give the van der Waals equation of state that works over the entire phase diagram, and it does not give the microscopic interpretation of the van der Waals parameters a and b. For this type of information, we need a molecular theory. Even so, it is remarkable to see how much we can learn from the simple assumption of Landau theory.

Note that the cubic term $\frac{1}{3}c(\delta v)^3$ does not play an important role in this argument, because it is small near the critical point. This term and higher-order terms would be important farther from the critical point—but we cannot rely on Landau theory far from the critical point, because the power series expansion is only justified for small δv, δp, and δT.

If the system is nonuniform, we can generalize Landau theory for the gas-liquid system just as we generalized Landau theory for the Ising model. The free energy density now acquires an extra term that represents the free energy cost of nonuniform $\delta v(\mathbf{r})$. To the lowest order, this cost can be written as

$$f_{\text{gradient}} = \frac{1}{2}K|\nabla(\delta v)|^2. \tag{4.29}$$

Hence, the full free energy density is

$$f = f_0 + a\delta v + \frac{1}{2}b(\delta v)^2 + \frac{1}{3}c(\delta v)^3 + \frac{1}{4}d(\delta v)^4 + \frac{1}{2}K|\nabla(\delta v)|^2, \tag{4.30}$$

and the integrated free energy is

$$F = \int d^3r \left[f_0 + a\delta v + \frac{1}{2}b(\delta v)^2 + \frac{1}{3}c(\delta v)^3 + \frac{1}{4}d(\delta v)^4 + \frac{1}{2}K|\nabla(\delta v)|^2 \right]. \tag{4.31}$$

This expression will be useful for the study of interfaces in Chap. 6.

4.3 General Order Parameters

In the previous two sections, we have developed Landau theories for two types of phase transitions: Ising and gas-liquid transitions. We have seen that these Landau theories are very similar. They both involve representing the free energy as a power series in the order parameter (M for Ising, δv for gas-liquid transition), and assuming that the coefficients in the series expansion are smooth functions of the thermodynamic variables (T and h for Ising, T and p for gas-liquid transition). To be sure, these two theories are not exactly the same because of a symmetry difference: The Ising model has an exact symmetry between $M > 0$ and $M < 0$, while the gas-liquid system does not have an exact symmetry between $\delta v > 0$ and $\delta v < 0$. For this reason, the gas-liquid system has odd as well as even terms in the power series for the free energy. Despite this difference, the systems are similar enough that they have phase diagrams with the same structure—a first-order transition terminating in a critical point—and the same critical exponents.

At this point, you might be wondering: Is that all there is? Are all phase transitions just like these two? Does every phase transition has the same type of Landau theory, and hence the same phase diagram and critical exponents?

The answer to these questions is no. To see the differences among phase transitions, we must consider the symmetries of the order parameters.

The Ising and gas-liquid transitions are so similar because their order parameters are both *real scalars*. They each have a magnitude and a sign, but no other directionality. Other systems may have order parameters with different mathematical structures:

- The order parameter may be a *vector*. One common example is a general ferromagnet—not just an Ising model, but a system where the magnetization can point in any direction in 3D space. In this case, the magnetic order parameter is a 3D vector.
- The order parameter may be a *complex number*. Common examples are superconductors and superfluids, where the order parameter $\psi = |\psi| e^{i\phi}$ has a magnitude and a phase. A related situation occurs for the layering order parameter of a smectic liquid crystal. In all these cases, the complex order parameter is equivalent to a two-component vector, because it can be broken up into real and imaginary parts.
- The order parameter may be a *tensor*. One common example is a nematic liquid crystal, which will be discussed later in Chap. 10.
- The system may have *several interacting order parameters*. For example, a smectic liquid crystal has an orientational order parameter and a layering order parameter, which interact with each other. In such cases, the phase diagram may be affected by the interaction of all the order parameters.

No matter what is the mathematical structure of the order parameter(s), the free energy always has the same mathematical structure: It is always a *real scalar*. Hence, whenever we construct a free energy as a power series, we must determine how to build a real scalar out of powers of the order parameter(s). From the perspective of Landau theory, we can say that the richness of phase diagrams arises from the many ways to combine interesting order parameters into a free energy.

4.4 Beyond Mean Field: Universality Classes

Landau theory is a form of mean-field theory, which expresses the free energy in terms of a small number of variables and then minimizes the free energy. It does not consider correlated fluctuations of many degrees of freedom at nearby positions. It gives the same critical exponents as the other forms of mean-field theory that we have studied.

I should briefly mention that there are other theories that go beyond the mean-field approximation; they do consider correlated fluctuations of many degrees of freedom at nearby positions. One very powerful approach to this challenge is the *renormalization group*, pioneered by Michael Fisher, Leo Kadanoff, and Kenneth Wilson, and recognized by the Nobel Prize to Wilson in 1982. This approach begins with a free energy density as in Landau theory, but it does not just minimize the free energy. Rather, it treats the partition function in a more subtle way, and it predicts critical exponents that are more precise than mean-field theory. This approach is beyond the scope of this book, and I cannot describe it further here. I only want to make sure you are aware that it exists!

Beyond mean-field theory, one of the main conclusions of statistical mechanics is that phase transitions can be classified into *universality classes*. Each universality class is characterized by a certain set of critical exponents. The universality class of a phase transition is determined by two quantities: the number of components in the order parameter (conventionally called n) and the dimensionality of space (called d). Hence, both the 3D Ising model and the 3D gas-liquid transition are in the universality class $n = 1$, $d = 3$. By comparison, if the magnetization can point in any direction in 3D space, the ferromagnet has $n = 3$, $d = 3$.

For the rest of this book, I will return to mean-field theory and will apply it to a range of problems in materials science. Although it does not give the precise critical exponents for phase transitions, it will prove to be remarkably useful in many situations!

Further Reading

Landau's view of statistical physics is presented in the following textbook:

1. L.D. Landau, E.M. Lifshitz, *Statistical Physics*, 3rd edn. Part 1 (Course of Theoretical Physics, vol 5). Translated by J.B. Sykes, M.J. Kearsley (Elsevier, 1980)

For a detailed discussion of critical phenomena, in mean-field theory and beyond, you should see:

2. S.-K. Ma, *Modern Theory of Critical Phenomena* (Westview, 2000)

Chapter 5
First Mathematical Interlude: Variational Calculus

Abstract Many problems in statistical mechanics require minimizing a free energy not just over one variable, or even several variables, but over a function. Variational calculus is the mathematical method for performing such minimizations. This chapter presents the mathematical technique of variational calculus, with examples in classical mechanics. In future chapters, this technique will be applied to the theory of soft materials.

In the preceding chapters, we have seen several examples where we need to minimize a free energy. In all of those examples, the free energy is a function of just one variable, the order parameter, which might be the magnetization M or the volume V. In these cases, we know how to do the minimization: We calculate the first derivative of the free energy with respect to the order parameter, set the first derivative equal to zero, and solve for the order parameter.

One simple generalization of those examples might be a problem where the free energy depends on *several* variables, perhaps several different order parameters. In this generalized problem, we would again know how to do the minimization: We would calculate the first *partial derivative* of the free energy with respect to each of the variables, set all of these partial derivatives equal to zero, and solve this system of equations for all of the variables. I will assume that you have already seen minimization problems like that in classes on multivariable calculus.

In statistical mechanics, we often need to deal with minimization problems that are yet more complicated: The order parameter itself may be a function of position, such as a local magnetization $M(\boldsymbol{r})$. In this case, the total free energy is the integral of a local free energy density, which depends on the local order parameter and its derivatives. In other words, the total free energy is a function of a function. We must then ask: What order parameter function gives the minimum free energy? Thus, the mathematical task is to minimize the free energy over all possible order parameter functions. This is a problem within the field of mathematics known as *variational calculus*, or *calculus of variations*.

In my experience, most students have not learned variational calculus at this point in their studies. For this reason, I will explain the basic mathematical concept here, so that we can use it in future chapters.

J.V. Selinger, *Introduction to the Theory of Soft Matter*,
Soft and Biological Matter, DOI 10.1007/978-3-319-21054-4_5

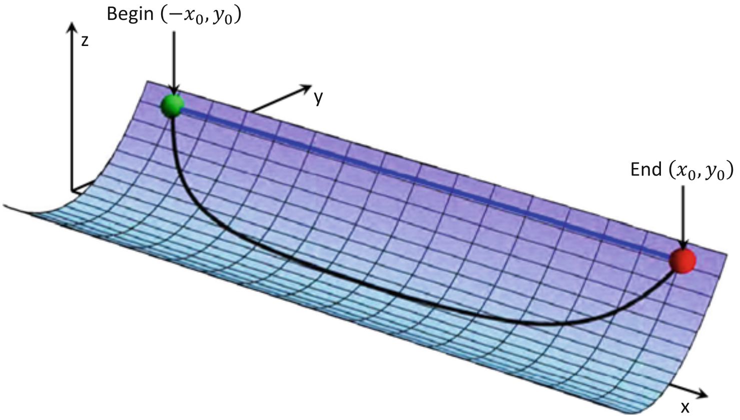


Fig. 5.1 Bobsled race on the side of a valley, as an example of a variational calculus problem (Interactive version at http://www.springer.com/cda/content/document/cda_downloaddocument/Selinger+Interactive+Figures.zip?SGWID=0-0-45-1509169-p177545420.)

5.1 Example: Bobsled Race

As an example of a variational calculus problem, let us consider an Olympic bobsled race on the side of a valley, as shown in Fig. 5.1. The race begins at the green point $(-x_0, y_0)$ and ends at the red point (x_0, y_0). As a racer, you need to choose your path $y(x)$. Clearly, the shortest *distance* is a straight line along the side of the valley, as shown by the blue line. However, what path is the shortest *time?* It might be advantageous for you to go down to the bottom of the valley in order to pick up speed, then proceed quickly along the bottom, and only go back up near the end of the race, as shown by the black line.

The first step in this problem is to derive an expression for the total travel time in terms of the path $y(x)$. The second step will be to minimize this expression over all possible paths.

To derive an expression for the total travel time, we need the bobsled speed at any point. Let us suppose that the valley has a parabolic shape

$$z = \frac{1}{2}ky^2. \tag{5.1}$$

To determine the speed v, we use the conservation of energy

$$\frac{1}{2}mv^2 + mgz = E_0, \tag{5.2}$$

where m is the mass, g is the gravitational acceleration, E_0 is the initial energy (kinetic plus potential), and we assume there is no drag on the bobsled. Hence, the speed at any point is

$$v = \left(\frac{2E_0}{m} - gky^2\right)^{1/2}. \tag{5.3}$$

Let us assume that the current potential energy is much less than the initial total energy, $\frac{1}{2}gky^2 \ll E_0$, so that the velocity can be approximated as

$$v = \left(\frac{2E_0}{m}\right)^{1/2}\left(1 - \frac{mgky^2}{4E_0}\right). \tag{5.4}$$

The total travel time T is now the integral of the time required for each element of the path, from beginning to end. We can write it as

$$T = \int_{\text{begin}}^{\text{end}} dt = \int_{\text{begin}}^{\text{end}} \frac{ds}{v}. \tag{5.5}$$

Here, ds is the element of arclength, which can be written as

$$\begin{aligned} ds =& \left(dx^2 + dy^2 + dz^2\right)^{1/2} \qquad (5.6)\\ =& dx\left(1 + \left(\frac{dy}{dx}\right)^2 + \left(\frac{dz}{dx}\right)^2\right)^{1/2}. \end{aligned}$$

Let us now assume that the horizontal slope is small, $dy/dx \ll 1$, and the vertical slope is even smaller, $dz/dx \ll dy/dx$. In this case, the element of arclength can be approximated as

$$ds = dx\left(1 + \frac{1}{2}\left(\frac{dy}{dx}\right)^2\right). \tag{5.7}$$

The total travel time then becomes

$$\begin{aligned} T =& \int_{-x_0}^{x_0} dx\left(1 + \frac{1}{2}\left(\frac{dy}{dx}\right)^2\right)\left(\frac{2E_0}{m}\right)^{-1/2}\left(1 - \frac{mgky^2}{4E_0}\right)^{-1}\\ =& \left(\frac{m}{2E_0}\right)^{1/2}\int_{-x_0}^{x_0} dx\left(1 + \frac{mgk}{4E_0}y^2 + \frac{1}{2}\left(\frac{dy}{dx}\right)^2\right). \qquad (5.8) \end{aligned}$$

Equation (5.8) for the total travel time is now our expression to minimize. This expression involves an integral over the whole path of the time required for each element of the path. It depends on the path $y(x)$ in two ways. First, the y^2 term favors keeping the path close to $y = 0$, on the bottom of the valley, where the speed is greatest. Second, the $(dy/dx)^2$ term favors keeping the path as straight as possible, in order to minimize the distance. These two terms compete with each other, and their competition determines the optimal path.

5.2 Minimization

How can we minimize an expression like Eq. (5.8) over paths $y(x)$? Let us reason by analogy with functions of one variable, or functions of several variables.

First, suppose we have a function of one variable $f(x)$. The basic idea is that we make a small change in $x \to x + \Delta x$, and then observe how f changes to $f(x) + \Delta f$. These changes are related by

$$f(x + \Delta x) = f(x) + \Delta f = f(x) + \frac{df}{dx}\Delta x + \text{higher-order terms.} \tag{5.9}$$

Canceling $f(x)$ from both sides of that equation gives

$$\Delta f = \frac{df}{dx}\Delta x + \text{higher-order terms.} \tag{5.10}$$

Hence, the first derivative df/dx expresses how much of a change in f results from a small change in x. If we are at the minimum (or maximum) of the function, then a small change in x should give *no change* in f, at linear order in Δx. It should only give a higher-order change in f that is quadratic in Δx. Thus, a condition for a minimum is

$$\frac{df}{dx} = 0. \tag{5.11}$$

This is a single equation that must be solved for the single variable x.

Now suppose we have a function of several variables $f(x_1, x_2, \ldots x_N)$. In this case, we can make small changes in *all* of the $x_i \to x_i + \Delta x_i$, and then observe how f changes to $f(x_1, x_2, \ldots x_N) + \Delta f$. These changes are related by

$$\begin{aligned} f(x_1 + \Delta x_1, x_2 + \Delta x_2, \ldots x_N + \Delta x_N) &= f(x_1, x_2, \ldots x_N) + \Delta f \\ &= f(x_1, x_2, \ldots x_N) + \sum_{i=1}^{N} \frac{\partial f}{\partial x_i}\Delta x_i + \text{higher-order terms.} \end{aligned} \tag{5.12}$$

Canceling $f(x_1, x_2, \ldots x_N)$ from both sides of the equation gives

$$\Delta f = \sum_{i=1}^{N} \frac{\partial f}{\partial x_i}\Delta x_i + \text{higher-order terms.} \tag{5.13}$$

Hence, the first partial derivative $\partial f/\partial x_i$ expresses how much of a change in f results from small changes in x_i. If we are at the minimum (or maximum or saddle point) of the function, then any set of small changes in the x_i should give no change in f, at linear order in Δx_i. It should only give a higher-order change in f that is quadratic in the Δx_i. Thus, a condition for a minimum is

$$\frac{\partial f}{\partial x_i} = 0 \tag{5.14}$$

for all i. This is a system of N equations that must be solved for the N variables x_i.

Finally, let us return to the variational calculus problem of the previous section. We have an expression for $T[y(x)]$ that depends on the function $y(x)$. We can make a small change in this function $y(x) \to y(x) + \Delta y(x)$ and then observe how T changes to $T[y(x)] + \Delta T$. These changes are related by

$$T[y(x) + \Delta y(x)] = T[y(x)] + \Delta T \tag{5.15}$$

$$= T[y(x)] + \int dx \frac{\delta T}{\delta y(x)} \Delta y(x) + \text{higher-order terms.} \tag{5.16}$$

Canceling $T[y(x)]$ from both sides of the equation gives

$$\Delta T = \int dx \frac{\delta T}{\delta y(x)} \Delta y(x) + \text{higher-order terms.} \tag{5.17}$$

Here, the quantity $\delta T / \delta y(x)$ is the *functional derivative*, which expresses how much of a change in T results from a small change in the function $y(x)$ at each point x. If we are at the minimum (or maximum or saddle point) of T, then small changes in $y(x)$ should give no change in T, at linear order in $\Delta y(x)$. It should only give a higher-order change in T that is quadratic in the $\Delta y(x)$. Thus, a condition for a minimum is

$$\frac{\delta T}{\delta y(x)} = 0 \tag{5.18}$$

for all x.

How do we calculate the functional derivative? Let us return to the expression of Eq. (5.8), and explicitly change $y(x) \to y(x) + \Delta y(x)$. This calculation gives

$$T[y(x) + \Delta y(x)] = \tag{5.19}$$
$$\left(\frac{m}{2E_0}\right)^{1/2} \int_{-x_0}^{x_0} dx \left(1 + \frac{mgk}{4E_0}(y + \Delta y)^2 + \frac{1}{2}\left(\frac{d(y + \Delta y)}{dx}\right)^2\right).$$

Now we expand this expression, and organize the terms based on the powers of Δy, to obtain

$$\begin{aligned}
&T[y(x)+\Delta y(x)] = \\
&\left(\frac{m}{2E_0}\right)^{1/2}\int_{-x_0}^{x_0} dx\left(1+\frac{mgk}{4E_0}y^2+\frac{1}{2}\left(\frac{dy}{dx}\right)^2\right). \\
&+\left(\frac{m}{2E_0}\right)^{1/2}\int_{-x_0}^{x_0} dx\left(\frac{mgk}{2E_0}y\Delta y+\frac{dy}{dx}\frac{d\Delta y}{dx}\right). \\
&+\left(\frac{m}{2E_0}\right)^{1/2}\int_{-x_0}^{x_0} dx\left(\frac{mgk}{4E_0}(\Delta y)^2+\frac{1}{2}\left(\frac{d\Delta y}{dx}\right)^2\right).
\end{aligned} \tag{5.20}$$

On the right-hand side of this equation, the first line is just the original $T[y(x)]$. The third line is higher-order terms, which are quadratic in Δy. Hence, the change ΔT is just

$$\Delta T=\left(\frac{m}{2E_0}\right)^{1/2}\int_{-x_0}^{x_0} dx\left(\frac{mgk}{2E_0}y\Delta y+\frac{dy}{dx}\frac{d\Delta y}{dx}\right)+\text{higher-order terms.} \tag{5.21}$$

Let us compare Eqs. (5.17) and (5.21). We are looking for an expression in the form of Eq. (5.17), an integral of something times $\Delta y(x)$, which would allow us to read off the functional derivative $\delta T/\delta y(x)$. Equation (5.21) is *almost* in the form that we want. The first term is the integral of something times $\Delta y(x)$, which is perfect. Unfortunately, the second term has a factor of $d\Delta y/dx$, rather than $\Delta y(x)$ itself. However, we can put it into the right form through integration by parts. As you recall, integration by parts involves using the identity

$$\int u dv = uv - \int v du. \tag{5.22}$$

We can apply this identity to the second term in Eq. (5.21), using

$$u=\frac{dy}{dx},\quad du=\frac{d^2y}{dx^2}dx,\quad v=\Delta y,\quad dv=\frac{d\Delta y}{dx}dx. \tag{5.23}$$

We then obtain

$$\int_{-x_0}^{x_0}\frac{dy}{dx}\frac{d\Delta y}{dx}dx=\left[\frac{dy}{dx}\Delta y\right]_{-x_0}^{x_0}-\int_{-x_0}^{x_0}\Delta y\frac{d^2y}{dx^2}dx. \tag{5.24}$$

Recall that we have two boundary conditions: The race begins at a fixed point $(-x_0, y_0)$ and ends at the fixed point (x_0, y_0). Equivalently, the function $y(x)$ must satisfy $y(-x_0)=y(x_0)=y_0$. The function $y(x)$ cannot have any variation at these two endpoints, but only in the interior. For this reason, the small change $\Delta y(x)$ must satisfy $\Delta y(-x_0)=\Delta y(x_0)=0$. Hence, on the right-hand side of Eq. (5.24), the first term must be zero, and we have only the second term. Putting this relation back into Eq. (5.21) then gives

$$\Delta T = \left(\frac{m}{2E_0}\right)^{1/2} \int_{-x_0}^{x_0} dx \left(\frac{mgk}{2E_0} y - \frac{d^2 y}{dx^2}\right) \Delta y(x) + \text{higher-order terms.} \tag{5.25}$$

Equation (5.25) is now in *exactly* the same form as Eq. (5.17). Hence, we can read off the functional derivative $\delta T/\delta y(x)$ as the coefficient of $\Delta y(x)$ in the integral:

$$\frac{\delta T}{\delta y(x)} = \left(\frac{m}{2E_0}\right)^{1/2} \left(\frac{mgk}{2E_0} y - \frac{d^2 y}{dx^2}\right). \tag{5.26}$$

We can now use this functional derivative to solve the minimization problem. The condition for a minimum is

$$\frac{\delta T}{\delta y(x)} = \left(\frac{m}{2E_0}\right)^{1/2} \left(\frac{mgk}{2E_0} y - \frac{d^2 y}{dx^2}\right) = 0, \tag{5.27}$$

which implies that

$$\frac{d^2 y}{dx^2} = \frac{mgk}{2E_0} y. \tag{5.28}$$

In general, in any variational calculus problem, the equation that the functional derivative equals zero is called the *Euler–Lagrange equation* for the problem. Note that the Euler–Lagrange equation is a differential equation for the function $y(x)$; we must solve this differential equation to find the minimum. In this problem, the solution of the differential equation is straightforward:

$$y(x) = C_1 e^{x/\xi} + C_2 e^{-x/\xi}, \tag{5.29}$$

where

$$\xi = \left(\frac{2E_0}{mgk}\right)^{1/2}, \tag{5.30}$$

and C_1 and C_2 are constants of integration. These constants can be determined by the boundary conditions, which give

$$C_1 = C_2 = \frac{y_0}{e^{x_0/\xi} + e^{-x_0/\xi}}, \tag{5.31}$$

and hence

$$y(x) = y_0 \frac{e^{x/\xi} + e^{-x/\xi}}{e^{x_0/\xi} + e^{-x_0/\xi}} = \frac{y_0 \cosh(x/\xi)}{\cosh(x_0/\xi)}. \tag{5.32}$$

This solution is plotted as the black line in Fig. 5.1.

The length scale ξ from Eq. (5.30) is an important feature of the solution. If ξ is small, the bobsled goes rapidly to a low elevation, near $y = 0$, in order to pick up speed. This behavior is favored by large values of m, g, and k, which cause gravitational effects to be important. By contrast, if ξ is large, the bobsled avoids changing its elevation rapidly, in order to minimize the length of the path. This behavior is favored by small values of m, g, and k, so that gravitational effects are not important compared with the large initial energy E_0. (The interactive version of Fig. 5.1 allows you to vary ξ and see the change in the path.)

By the way, some students are familiar with the Euler–Lagrange equation in the form

$$\frac{\partial \mathcal{T}}{\partial y} - \frac{d}{dx}\left(\frac{\partial \mathcal{T}}{\partial y'}\right) = 0, \tag{5.33}$$

where $y' = dy/dx$ and $\mathcal{T}$ is the integrand of Eq. (5.8). You can verify explicitly that this equation is equivalent to the Euler–Lagrange equation derived here. In general, Eq. (5.33) works as long as the integral to be minimized depends only on the function $y(x)$ and its first derivative (which is usually true). It needs to be generalized if the integral to be minimized depends on higher derivatives (which happens on rare occasions). Personally, I find this equation to be somewhat unintuitive, so I prefer to derive the functional derivative and explicitly set it equal to zero.

Problem: Repeat the bobsled problem on the side of a hill where $z = by$, instead of a parabolic valley. Find the path $y(x)$ that minimizes the travel time in this geometry.

Solution: On this hillside, the conservation of energy gives the bobsled speed

$$v = \left(\frac{2E_0}{m} - 2gby\right)^{1/2} \approx \left(\frac{2E_0}{m}\right)^{1/2}\left(1 - \frac{mgby}{2E_0}\right). \tag{5.34}$$

Hence, the total travel time is the integral

$$T = \left(\frac{m}{2E_0}\right)^{1/2} \int_{-x_0}^{x_0} dx \left(1 + \frac{mgb}{2E_0} y + \frac{1}{2}\left(\frac{dy}{dx}\right)^2\right). \tag{5.35}$$

By setting the functional derivative $\delta T/\delta y(x) = 0$, we obtain the Euler–Lagrange equation:

$$\frac{d^2 y}{dx^2} = \frac{mgb}{2E_0}. \tag{5.36}$$

The solution of this differential equation, subject to the boundary conditions that $y = y_0$ at $x = \pm x_0$, is

$$y(x) = y_0 + \frac{mgb}{4E_0}\left(x^2 - x_0^2\right). \tag{5.37}$$

Therefore, the optimal path on the hillside is a parabola, in contrast with Fig. 5.1, where the optimal path in the valley is a hyperbolic cosine. This difference is physically reasonable: In a valley, the optimal path is approximately straight along the valley floor. On a hillside, there is no floor, so the optimal path is always curved with the same second derivative.

5.3 Example: Guitar String

For a second example, consider the physics of a guitar string that is stretched between two posts at $x = 0$ and L, as shown in Fig. 5.2. The string is under some tension T (not to be confused with the travel time in the previous section!). Because of the tension, it has an elastic energy proportional to its arclength

$$U = T \cdot \text{arclength} = T \int_{\text{begin}}^{\text{end}} ds = T \int_{\text{begin}}^{\text{end}} \left(dx^2 + dy^2\right)^{1/2} \tag{5.38}$$

$$= T \int_0^L dx \left(1 + \left(\frac{dy}{dx}\right)^2\right)^{1/2}. \tag{5.39}$$

In the limit of small deviations from a straight horizontal string, $dy/dx \ll 1$, the energy can be approximated by

$$U = T \int_0^L dx \left(1 + \frac{1}{2}\left(\frac{dy}{dx}\right)^2\right) = TL + \frac{1}{2} T \int_0^L dx \left(\frac{dy}{dx}\right)^2. \tag{5.40}$$

To minimize the elastic energy, we must calculate the functional derivative of U with respect to $y(x)$, and set it equal to zero. Following the procedure of the previous section, the functional derivative is

$$\frac{\delta U}{\delta y(x)} = -T \frac{d^2 y}{dx^2}. \tag{5.41}$$

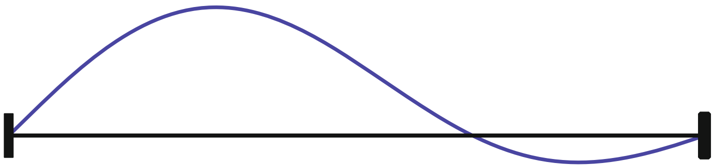


Fig. 5.2 Guitar string stretched between two posts at $x = 0$ and L

Hence, the Euler–Lagrange equation for this problem is just

$$\frac{d^2y}{dx^2} = 0. \tag{5.42}$$

This equation implies that the lowest-energy shape of the string is a straight line from the first post to the second post. This result is not a big surprise!

We can now see a more interesting feature of the problem: The functional derivative is not only useful for deriving the lowest-energy shape of the string; it is also useful for deriving the equation of motion. As you recall, the standard equation of motion for a particle is Newton's second law:

$$ma = F. \tag{5.43}$$

If the force F is derived from a potential energy U, then this equation becomes

$$m\frac{d^2y}{dt^2} = -\frac{\partial U}{\partial y} \tag{5.44}$$

for the coordinate y. For a guitar string described by the continuous function $y(x, t)$, the analogous equation is

$$\rho\frac{\partial^2 y}{\partial t^2} = -\frac{\delta U}{\delta y(x, t)}, \tag{5.45}$$

where ρ is the mass density per unit length of the string. Using our result of Eq. (5.41) for the functional derivative, the equation of motion becomes

$$\rho\frac{\partial^2 y}{\partial t^2} = T\frac{\partial^2 y}{\partial x^2}. \tag{5.46}$$

From your previous physics courses, you may recognize this equation as the *wave equation!* Hence, the functional derivative is providing us with a generalization of the concept of force, which allows us to derive the dynamics of waves on the string.

Further Reading

Variational calculus is one of many mathematical methods presented in the following textbooks:

1. G.B. Arfken, H.J. Weber, F.E. Harris, *Mathematical Methods for Physicists: A Comprehensive Guide*, 7th edn. (Elsevier, 2013)
2. M. Stone, P. Goldbart, *Mathematics for Physics: A Guided Tour for Graduate Students* (Cambridge, 2009)

A pre-publication version of the second book is available from the author's website http://www.physics.gatech.edu/~pgoldbart6/PG_MS_MfP.htm.

Chapter 6
Field Theory for Nonuniform Systems

Abstract Field theory is a powerful approach for understanding nonuniform systems in statistical mechanics, as well as other areas of physics. This chapter introduces the concept of a order parameter field, and then uses that concept to determine how an order parameter varies near an aligning wall, or at the interface between two ordered phases. The latter calculation leads to the concept of surface tension as the excess free energy associated with an interface.

In this chapter, I will introduce the concept of *field theory,* as applied to statistical mechanics. I will then use this concept to predict the behavior of nonuniform systems—either magnets or liquids and gases—near a critical point. In the end, this calculation will lead to the concept of *surface tension*.

6.1 What is a Field?

In any theoretical physics problem, we have to consider: What type of objects are we modeling? What are the fundamental degrees of freedom in our theory?

In some cases, the fundamental objects are *particles,* like molecules or bobsleds. Particles are individual objects, which each have a small number of degrees of freedom. Typically, they have six degrees of freedom: three components of position and three components of momentum. In some cases, they might have additional internal degrees of freedom, such as spin, but it is still a modest number. For particles, we want to predict how their coordinates $\boldsymbol{r}(t)$ and $\boldsymbol{p}(t)$ evolve as functions of time. If there are only a few particles (like one bobsled), we might make these predictions exactly. If there are many particles (like 10^{23} molecules), we might only be able to make predictions on a statistical basis.

In other cases, the fundamental objects are *fields.* A field just means a function of position, or of position and time. Most students first encounter fields in the context of electric and magnetic fields, $\boldsymbol{E}(\boldsymbol{r}, t)$ and $\boldsymbol{B}(\boldsymbol{r}, t)$. The important point to notice about fields is that they have infinitely many degrees of freedom. For example, the electric field $\boldsymbol{E}(\boldsymbol{r}, t)$ has three degrees of freedom at every point in space. For this reason, we cannot just ask how some particular degree of freedom evolves in time. Rather,

J.V. Selinger, *Introduction to the Theory of Soft Matter*,
Soft and Biological Matter, DOI 10.1007/978-3-319-21054-4_6

we must ask how the field $\boldsymbol{E}(\boldsymbol{r}, t)$ evolves as a function of both position and time. For the example of electricity and magnetism, this evolution is given by Maxwell's equations.

Apart from electricity and magnetism, the concept of a field also occurs in mechanics. The height of a guitar string $y(x, t)$ is a simple one-dimensional example, as discussed in the last chapter. The height of a drum $z(x, y, t)$ is a related two-dimensional example. The concept of a field also occurs frequently in fluid dynamics: People study the pressure field $p(\boldsymbol{r}, t)$, the density field $\rho(\boldsymbol{r}, t)$, and the velocity field $\boldsymbol{v}(\boldsymbol{r}, t)$.

In statistical physics, we consider an *order parameter field.* An order parameter does not have to be uniform; rather, it can vary as a function of position (or position and time). In this case, the order parameter field describes the statistical ordering in a small local region. For example, in the Ising model, the magnetic order parameter field $M(\boldsymbol{r})$ describes the average order of Ising spins near the position $\boldsymbol{r}$. In a liquid-gas problem, the local density $\rho(\boldsymbol{r})$ describes the average density of molecules near the position $\boldsymbol{r}$.

You should notice that the concept of an order parameter field is an example of *coarse-graining,* as discussed at the end of Chap. 1. At some microscopic level, an Ising magnet really is made of individual spins, and a liquid or gas really is made of molecules. These spins or molecules are particles. When we make a coarse-grained theory, we average over the spins or molecules and describe the material in a less detailed way. We then speak as if the order parameter field were the fundamental object in the theory, and neglect the microscopic spins or molecules.

Figure 6.1 shows an example of coarse-graining the Ising model. In part (a), we see an Ising model on a 2D square lattice of 200 × 120 spins. Each dark-colored square represents a down spin, and each light-colored square represents an up spin. Clearly, this system is nonuniform: On the left side, the spins are mostly pointing down, and on the right side, the spins are mostly pointing up. Hence, the magnetic order parameter M must be a function of x. In other words, the system has an order parameter field $M(x)$. The field is plotted in part (b), which shows that it increases linearly from the left to the right side.

Now you might ask: What is the right length scale for the averaging? In the Ising model, should $M(\boldsymbol{r})$ be averaged over a local group of 10 spins, or 100 spins, or some other number? In a molecular system, should the local density be averaged over a volume of $(1\,\text{nm})^3$, or $(10\,\text{nm})^3$, or $(100\,\text{nm})^3$? In general, there are two rules:

1. The averaging length scale should be much *bigger* than the size of a spin, or molecule, or other particle. If this constraint is satisfied, then it makes sense to describe the system as a field, rather than as individual particles.
2. The averaging length scale should be much *smaller* than the length scale over which the physical properties vary. (For example, in Fig. 6.1, the averaging length scale should be much smaller than the length scale over which the magnetic order parameter varies.) If this constraint is satisfied, then it makes sense to consider all the particles in the averaging volume as if they are part of the same statistical ensemble.

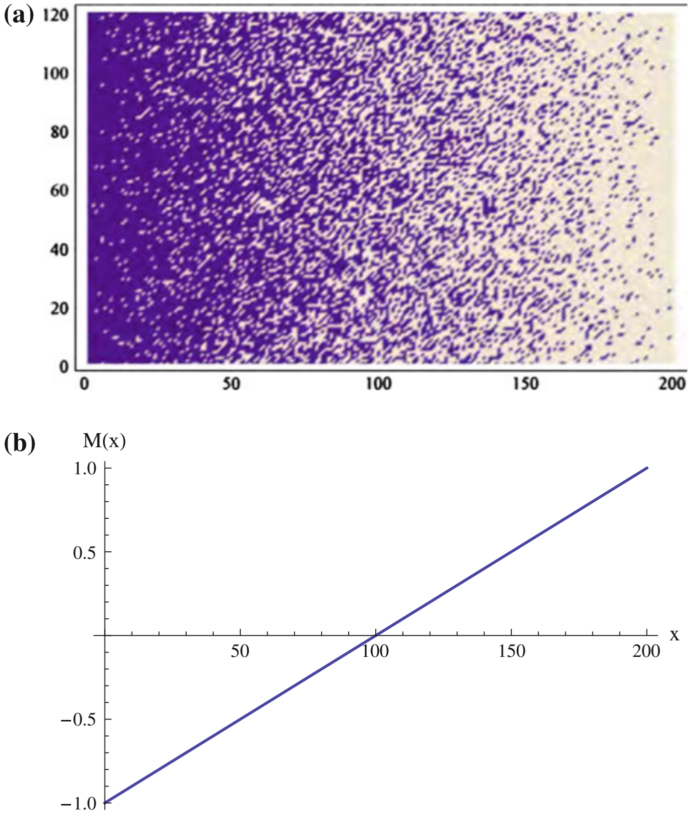


Fig. 6.1 Example of a nonuniform Ising model. In part **a**, each *dark-colored square* represents a down spin, and each *light-colored square* represents an up spin. In part **b**, the coarse-grained order parameter $M(x)$ is plotted as a function of x

Is it possible to find some averaging length scale that satisfies both of these constraints? In most experimental systems, yes! Molecules generally have a length scale of about 1 nm, and physical properties typically vary over a length scale of 1 µm (the wavelength of light, the size of a biological cell, the size of a liquid-crystal cell). There is a lot of room between 1 nm and 1 µm. As long as we stay within this range, it generally does not matter what length scale we use for averaging. For this reason, it is possible to describe physical systems using field theory, and it is usually not necessary to discuss the averaging length scale explicitly. You should just remember: *Field theory works because molecules are very, very small!*

By the way, the field theory that we are discussing here does not involve quantum mechanics, so we might call it *classical field theory*. By comparison, theoretical high-energy physics is based on a quantum-mechanical version of this approach, which is called *quantum field theory*. There are many interesting analogies between classical and quantum field theory, but that topic is beyond the scope of this book.

6.2 Nonuniform System in a Disordered Phase

For a first example, consider an Ising magnet with a wall. To be specific, suppose we have a wall at $x = 0$, and then Ising spins from $x = 0$ to ∞. This type of system is called *semi-infinite,* because it is infinite for positive x but cut off by the wall for negative x. The region of the sample far from the wall is called the *bulk,* and it should have the same properties that we calculated back in Chap. 2. However, the region of the sample near the wall is somewhat different just because of the presence of the wall, so it may have some different properties. In particular, the magnetic order parameter field $M(x)$ may be different near the wall than in the bulk.

Suppose we are in the disordered phase, with $T > T_C$, and there is no magnetic field. In this case, the bulk will certainly have $M = 0$. Now suppose that the wall tends to align the spins with perfect order pointing up, so that $M = 1$ at $x = 0$. (Why would the wall do that? Well, perhaps it has a very strong local magnetic field right at $x = 0$, although there is no magnetic field in the bulk.)

In this problem, we must ask: What is the order parameter field $M(x)$ as a function of distance away from the wall? How far does the magnetic order penetrate into the sample?

To address this problem, let us use Landau theory for the Ising model, as in Chap. 4. From Eq. (4.21), with field $h = 0$, the total free energy is an integral over position

$$F = \int d^3r \left[f_0 + \frac{1}{2}a'(T - T_C)M(\boldsymbol{r})^2 + \frac{1}{4}b_0 M(\boldsymbol{r})^4 + \frac{1}{2}K|\nabla M|^2 \right]. \tag{6.1}$$

We do not expect M to depend on y or z, but only on x, the distance away from the wall. Hence, we can replace the y and z integrals by factors of L_y and L_z, the system size in those two directions, to obtain

$$F = L_x L_y \int_0^\infty dx \left[f_0 + \frac{1}{2}a'(T - T_C)M(x)^2 + \frac{1}{4}b_0 M(x)^4 + \frac{1}{2}K\left(\frac{dM}{dx}\right)^2 \right]. \tag{6.2}$$

Because $T > T_C$, we expect $M(x)$ to be very small almost everywhere. Hence, the quartic term is very small in comparison with the quadratic term, so we can neglect the quartic term to obtain

$$F = L_x L_y \int_0^\infty dx \left[f_0 + \frac{1}{2}a'(T - T_C)M(x)^2 + \frac{1}{2}K\left(\frac{dM}{dx}\right)^2 \right]. \tag{6.3}$$

The question is now: What order parameter field $M(x)$ minimizes the free energy of Eq. (6.3)? This is a variational calculus problem, and we can solve it using the methods of Chap. 5. First, we must take the functional derivative of F with respect to $M(x)$. We can see that the free energy integral of Eq. (6.3) has the same types of terms as the travel time integral of Eq. (5.8) for the bobsled problem, so we can calculate the functional derivative in exactly the same way. The result is

$$\frac{\delta F}{\delta M(x)} = L_x L_y \left[a'(T - T_C)M(x) - K\frac{d^2M}{dx^2} \right]. \tag{6.4}$$

Hence, the Euler–Lagrange equation for this problem is

$$\frac{\delta F}{\delta M(x)} = 0, \tag{6.5}$$

which implies that

$$K\frac{d^2M}{dx^2} = a'(T - T_C)M(x). \tag{6.6}$$

The solution of this differential equation is

$$M(x) = C_1 e^{x/\xi} + C_2 e^{-x/\xi}, \tag{6.7}$$

where

$$\xi = \left[\frac{K}{a'(T - T_C)} \right]^{1/2}, \tag{6.8}$$

and C_1 and C_2 are constants of integration. These constants can be determined by the boundary conditions. The requirement that $M \to 0$ as $x \to \infty$ gives $C_1 = 0$, and the requirement that $M = 1$ at $x = 0$ gives $C_2 = 1$. Hence, our solution is

$$M(x) = e^{-x/\xi}. \tag{6.9}$$

Figure 6.2 shows a plot of the solution for $M(x)$. As required, it begins at $M = 1$ at the wall, and then gradually drops off to $M = 0$ in the bulk. This plot can be regarded as a *correlation function,* which shows how well the spins at x are correlated with the fixed spins at the wall. The length scale ξ is the *correlation length,* which shows what distance is required for the correlation function to decay to $1/e$ of its value at the wall.

The correlation length is controlled by the competition between two terms in the free energy of Eq. (6.3). The M^2 term favors the bulk state $M = 0$ and gives a free energy penalty for any deviations from $M = 0$. Hence, it favors small ξ, so that M can rapidly approach the favored value of 0. By contrast, the $(dM/dx)^2$ term favors uniform M and gives a free energy penalty for any gradients in M. Hence, it favors large ξ, so that gradients in M are small. The competition between these two terms gives Eq. (6.8) for the correlation length.

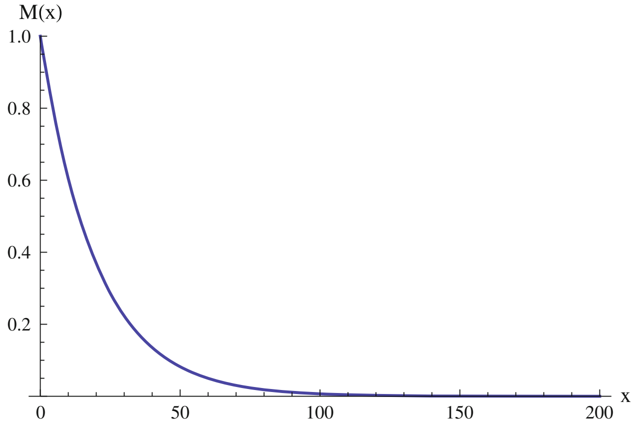


Fig. 6.2 Ising model in the disordered phase near an aligning wall. The order parameter profile $M(x)$ is shown as a function of x (Interactive version at http://www.springer.com/cda/content/document/cda_downloaddocument/Selinger+Interactive+Figures.zip?SGWID=0-0-45-1509169-p177545420.)

We can see that the correlation length depends on temperature in an interesting way. At high temperature, the correlation length is very small, so that the magnetic order drops away rapidly as a function of distance from the wall. As T decreases, the correlation length increases, so that the magnetic order persists over a longer length scale. As T approaches T_C, the correlation length diverges, following the scaling relation

$$\xi \propto |T - T_C|^{-\nu}, \tag{6.10}$$

where ν is another critical exponent. Equation (6.8) shows that the mean-field prediction for this critical exponent is $\nu = \frac{1}{2}$.

Because the liquid-gas system is described by the same type of Landau theory as the Ising magnet, we can expect the same type of nonuniform order parameter field to occur in the liquid-gas system also. For an analogous problem, suppose we have a supercritical fluid for $T > T_C$, and this fluid has a wall that favors high density (i.e., it favors the liquid phase). Based on the theory in this section, we expect to see a density profile similar to Fig. 6.2. Near the wall, the density is elevated. As we move away from the wall, the density drops down to its value in the bulk supercritical fluid. The excess density decays exponentially, with a correlation length ξ. This correlation length is small at high temperature, and it increases as the temperature decreases toward T_C, following the scaling relation of Eq. (6.10).

6.3 Interface Between Ordered Phases

As a second example, consider an Ising magnet at $T < T_C$, with no magnetic field. Because the temperature is below T_C, the system must have some spontaneous magnetic order $M = \pm M_0$, which may be either pointing down or pointing up. Suppose it is pointing down on one side of the system and pointing up on the other side. To be specific, suppose we have an infinite system with $M \to -M_0$ as $x \to -\infty$, and $M \to +M_0$ as $x \to +\infty$. Somewhere between these limits, the system must form an interface that separates the spin-down side from the spin-up side. What can we say about the width and the energy of this interface?

To answer this question, we must determine what order parameter field $M(x)$ minimizes the free energy. For this reason, we return to the free energy integral of Eq. (6.2):

$$F = L_x L_y \int_{-\infty}^{\infty} dx \left[f_0 + \frac{1}{2}a'(T - T_C)M(x)^2 + \frac{1}{4}b_0 M(x)^4 + \frac{1}{2}K\left(\frac{dM}{dx}\right)^2 \right]. \tag{6.11}$$

Now, because $T < T_C$, the coefficient of the quadratic term is negative. Hence, we cannot neglect the quartic term in the free energy; we need the quartic term for thermodynamic stability, so that the free energy will not go to $-\infty$ as M becomes large. We must minimize the free energy integral with the quadratic term, the quartic term, and the gradient term.

As a first step, consider the ordered domains away from the interface. Here, the gradient term is negligible and $M = \pm M_0$. In this case, the free energy density is

$$f = f_0 + \frac{1}{2}a'(T - T_C)M_0^2 + \frac{1}{4}b_0 M_0^4. \tag{6.12}$$

Minimizing this expression over M_0 gives

$$M_0 = \left[\frac{a'(T_C - T)}{b_0}\right]^{1/2}. \tag{6.13}$$

Hence, M_0 is not an independent parameter in the problem; it is controlled by the coefficients in the free energy.

Next, we can do the full minimization over the order parameter field $M(x)$. The functional derivative of the full free energy (including the quartic term) is

$$\frac{\delta F}{\delta M(x)} = L_x L_y \left[a'(T - T_C)M(x) + b_0 M(x)^3 - K\frac{d^2 M}{dx^2} \right]. \tag{6.14}$$

Hence, the Euler–Lagrange equation for this problem is

$$K\frac{d^2M}{dx^2} = a'(T - T_C)M(x) + b_0 M(x)^3. \tag{6.15}$$

We need to solve this equation for $M(x)$, with the boundary conditions that $M \to -M_0$ as $x \to -\infty$, and $M \to +M_0$ as $x \to +\infty$. This differential equation is much more challenging to solve than Eq. (6.6) because it is not linear in $M(x)$. Nevertheless, it has an exact solution. By direct substitution, you can verify that the differential equation and boundary conditions are exactly satisfied by

$$M(x) = M_0 \tanh\left(\frac{x - x_0}{\xi}\right), \tag{6.16}$$

where

$$\xi = \left[\frac{2K}{a'(T_C - T)}\right]^{1/2} \tag{6.17}$$

and x_0 is any constant position.

Figure 6.3 shows an example of the solution for $M(x)$. In this solution, x_0 is the center of the interface. On one side of the interface, M goes to $-M_0$, and on the other side, it goes to $+M_0$. The parameter ξ is a correlation length, which shows the thickness of the interface, i.e., the characteristic length scale associated with the change from $-M_0$ to $+M_0$. From Eq. (6.17), we see that ξ has an interesting dependence on temperature. At very low temperature $T \ll T_C$, the correlation length

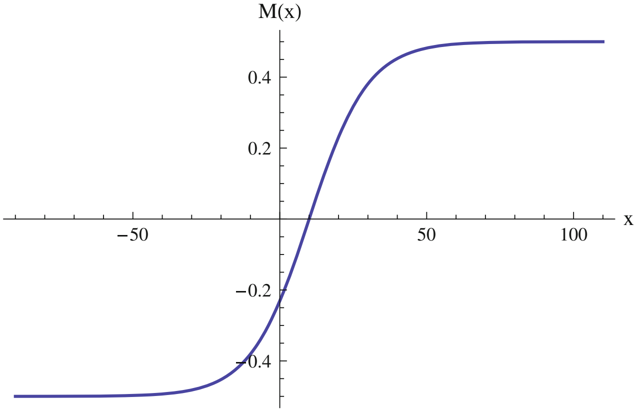


Fig. 6.3 Ising model in the ordered phase, showing the profile $M(x)$ through an interface between spin-down and spin-up sides (Interactive version at http://www.springer.com/cda/content/document/cda_downloaddocument/Selinger+Interactive+Figures.zip?SGWID=0-0-45-1509169-p177545420.)

is small, and hence the interface is very sharp. As T increases, ξ becomes larger, and hence the interface becomes broader. As T approaches T_C from below, ξ diverges following the scaling relation of Eq. (6.10), with the same critical exponent $\nu = \frac{1}{2}$ that we calculated in the previous section, for T approaching T_C from above.

In addition to calculating the order parameter $M(x)$, we can also calculate the excess free energy associated with the interface. Within a single ordered domain of $M = \pm M_0$, we have the minimum free energy density

$$f_{\min} = f_0 + \frac{1}{2}a'(T - T_C)M_0^2 + \frac{1}{4}b_0 M_0^4 = f_0 - \frac{a'^2(T_C - T)^2}{4b_0}. \tag{6.18}$$

Within the interface, where $M(x) \neq \pm M_0$, we have the excess free energy density

$$\begin{aligned}\Delta f(x) =& f(x) - f_{\min}\\ =&\frac{1}{2}a'(T - T_C)M(x)^2 + \frac{1}{4}b_0 M(x)^4 + \frac{1}{2}K\left(\frac{dM}{dx}\right)^2 + \frac{a'^2(T_C - T)^2}{4b_0}\\ =&\frac{a'^2(T_C - T)^2}{2b_0}\operatorname{sech}^4\left(\frac{x - x_0}{\xi}\right).\end{aligned} \tag{6.19}$$

Hence, the total excess free energy associated with the interface, compared with a single domain of $M = \pm M_0$, is the integral

$$\begin{aligned}\Delta F =& L_x L_y \int_{-\infty}^{\infty} dx\,\Delta f(x) = L_x L_y \frac{a'^2(T_C - T)^2}{2b_0}\int_{-\infty}^{\infty} dx\,\operatorname{sech}^4\left(\frac{x - x_0}{\xi}\right)\\ =& L_x L_y \frac{a'^2(T_C - T)^2}{2b_0}\frac{4\xi}{3} = L_x L_y \frac{2^{3/2}K^{1/2}a'^{3/2}(T_C - T)^{3/2}}{3b_0}.\end{aligned} \tag{6.20}$$

This equation shows that the excess free energy of the interface is some coefficient times the interfacial area $L_x L_y$. This coefficient is exactly what we mean by a *surface tension,* a free energy per unit area associated with the interface. Hence, we see that the surface tension of the interface between spin-down and spin-up sides is

$$\sigma = \frac{2^{3/2}K^{1/2}a'^{3/2}(T_C - T)^{3/2}}{3b_0}. \tag{6.21}$$

This surface tension goes to zero as $T \to T_C$, because the distinction between spin-down and spin-up phases vanishes at T_C.

As a matter of terminology, this type of variation in $M(\boldsymbol{r})$ is called a *soliton.* A soliton is a remarkable type of distortion, which has a characteristic width, a characteristic change in the order parameter, and a characteristic free energy per unit area. It cannot smooth out or break up into smaller distortions in $M(\boldsymbol{r})$. Physically, it occurs because the system has two stable states (spin-down and spin-up), and the soliton takes it from one stable state to the other. Mathematically, it occurs because the

Euler–Lagrange equation of Eq. (6.15) is a nonlinear equation. Solitons are generally a phenomenon of nonlinear waves, which do not satisfy the superposition principle of linear waves.

Like the problem in the previous section, this problem also has an analog in the liquid-gas system. In this case, the analogous system is at $T < T_C$, so that the free energy has two distinct minima at different densities, or different volumes per molecule, corresponding to the liquid and gas phases. The system can then have a nonuniform state with liquid on one side and gas on the other. This state can be described by a density field $\rho(x)$, or a volume per molecule field $v(x)$, with $v(x) \to v_L$ as $x \to -\infty$ and $v(x) \to v_G$ as $x \to +\infty$. Our goal is then to find the function $v(x)$ that minimizes the free energy integral.

Because the liquid-gas system has the same Landau theory as the Ising magnet, the problem of finding $v(x)$ is analogous to finding the Ising order parameter $M(x)$. The solution must have the same form as $M(x)$ shown in Fig. 6.3. Here, $v(x)$ has an interface, with v_L on one side and v_G on the other. The interface has a width ξ and a surface tension σ. At low temperature, when the liquid and gas phases are very different, the interface is very sharp (ξ is small), and the surface tension is high. As T approaches T_C, the interface width increases and the surface tension decreases. At the critical point, where the liquid and gas become identical, the interface width goes to infinity and the surface tension goes to zero. This analysis justifies Fig. 3.11, at the end of Chap. 3, which shows that the liquid-gas interface becomes fuzzy as the system approaches the critical point.

Further Reading

Most physics students learn about field theory in the context of beaded strings in classical mechanics, or electric and magnetic fields in electrodynamics. Engineering students often learn about stress and strain fields in solids, or pressure, density, and velocity fields in fluids. The concept of order parameter fields in statistical mechanics is not as well known as those other examples. For a careful discussion of this concept, I recommend the pair of advanced textbooks:

1. M. Kardar, *Statistical Physics of Particles* (Cambridge, 2007)
2. M. Kardar, *Statistical Physics of Fields* (Cambridge, 2007)

Chapter 7
Dynamics of Phase Transitions

Abstract This chapter considers how a system can evolve in time from one phase to another and presents the theory of nucleation and growth. This theory uses the concept of surface tension, developed in the last chapter, and provides the basis for understanding glasses in the next chapter.

In our study of phase transitions so far, we have assumed that a system is always in the equilibrium state, at the minimum of the free energy. This is generally true if we wait long enough—but what happens if we do not wait long enough? In this chapter, we will consider the dynamic process of phase transitions. We will see that the dynamic process of nucleation and growth involves surface tension, as discussed in the last chapter.

7.1 Nucleation and Growth

For a first example, consider an Ising magnet at low temperature $T < T_C$, under a negative applied field $h < 0$. Because of the negative field, the equilibrium magnetic order parameter is also negative $M < 0$. At some time, we suddenly change the field from negative to positive, as shown in the phase diagram of Fig. 7.1. The new equilibrium state then has $M > 0$. The question is: How does the system evolve from $M < 0$ to $M > 0$?

To understand this evolution, look at the free energy plot *after* the field change, as shown in Fig. 7.2. This plot has two minima, which are at $M = \pm 0.85$. Before the field change, the state with $M = -0.85$ was the absolute minimum, so the system was in that state. After the field change, the state with $M = -0.85$ is no longer the absolute minimum, but the system is still temporarily in that state. We say that the state with $M = -0.85$ is *metastable*, meaning that it is a local minimum but not the absolute minimum. By comparison, the state with $M = 0.85$ is *stable*, meaning that it is the absolute minimum.

If we begin with the system at $M = -0.85$, any *small* changes in the magnetic order will increase the free energy. As a result, if any small changes occur at random, they will tend to go away, and the system will return to the metastable state.

J.V. Selinger, *Introduction to the Theory of Soft Matter*,
Soft and Biological Matter, DOI 10.1007/978-3-319-21054-4_7

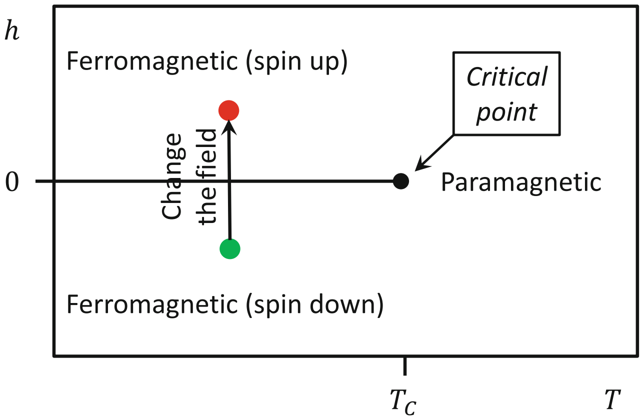


Fig. 7.1 Phase diagram of the Ising model, showing a dynamic change in the field from $h < 0$ to $h > 0$

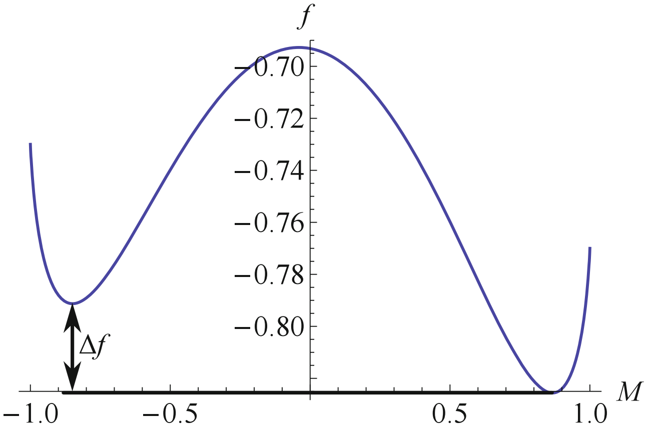


Fig. 7.2 Plot of the free energy density of the Ising model as a function of M, for $T < T_C$ and $h > 0$, showing the metastable state at $M = -0.85$ and the stable state at $M = 0.85$

The only way for the system to get out of its metastable state is if some *large* changes occur at random. One example of a large change is shown in Fig. 7.3. Here, most of the system is in the old state with the spins mostly pointing down. However, within the circle, the system has formed a cluster of spins that are mostly pointing up. This cluster is called a *nucleus* or *droplet* or *bubble* of the new state with $M = 0.85$, and the process of forming such a cluster is called *nucleation*.

In the language of order parameter fields, as in the last chapter, we would say that the nucleus is a place where $M(\boldsymbol{r}) \approx M_0$, which is the value in the new stable phase. By contrast, the bulk of the magnet has $M(\boldsymbol{r}) \approx -M_0$, which is the value in the old metastable phase. The boundary between these regions is an interface, similar to the interface discussed in Sect. 6.3. (It is not exactly the same because the states on the two sides of the interface do not have exactly the same free energy density, and

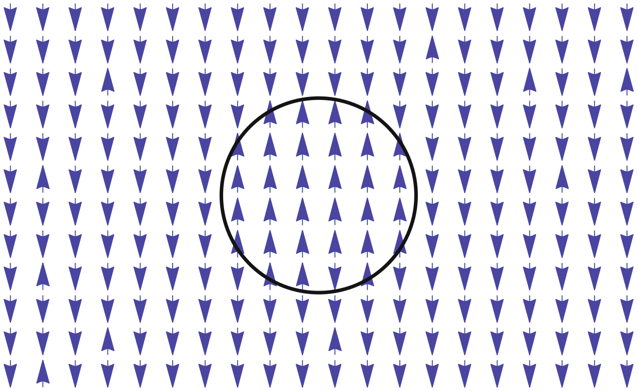


Fig. 7.3 Nucleus of the *spin-up* phase (with $M = 0.85$) forming within the *spin-down* phase ($M = -0.85$)

because the interface is not flat.) As in Sect. 6.3, the interface has some width ξ; it is not perfectly sharp. It also has some surface tension σ, an energy per unit area.

Once a nucleus of the new phase forms, the question is whether it will grow or shrink. To answer this question, consider a 3D spherical nucleus of radius r. We must calculate the free energy of this nucleus, compared with the bulk metastable state, as a function of r. This free energy has two contributions. First, there is a difference in free energy density between the stable spin-up state and the metastable spin-down state, which is shown as Δf in Fig. 7.2. For the whole nucleus with radius r, the difference in free energy density is multiplied by the volume of the nucleus to give

$$\Delta F_{\text{volume}} = -\frac{4}{3}\pi r^3 \Delta f. \tag{7.1}$$

Here, the minus sign shows that this contribution to the free energy is negative, because the stable state has a lower free energy density than the metastable state, taking Δf as a positive parameter. Second, we must also include the surface free energy, which is the surface tension σ times the surface area of the nucleus:

$$\Delta F_{\text{surface}} = 4\pi r^2 \sigma. \tag{7.2}$$

This contribution is positive, because the surface tension costs energy compared with the uniform metastable state. Putting these two pieces together, we obtain the total free energy of the nucleus:

$$\Delta F_{\text{nucleus}} = -\frac{4}{3}\pi r^3 \Delta f + 4\pi r^2 \sigma. \tag{7.3}$$

This expression for the free energy of a nucleus is plotted in Fig. 7.4. For small r, it is dominated by the positive surface term, and for large r, it is dominated by the negative volume term. As a result, it has a maximum at the *critical radius*

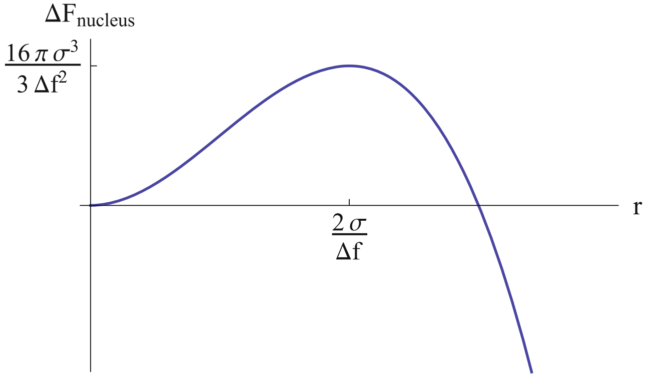


Fig. 7.4 Free energy of a 3D spherical nucleus, compared with the uniform metastable background, as a function of radius

$$r_C = \frac{2\sigma}{\Delta f}, \tag{7.4}$$

and the free energy at the critical radius is

$$\Delta F_{\text{nucleus}}(r_C) = \frac{16\pi\sigma^3}{3\Delta f^2}. \tag{7.5}$$

From this plot, we can see how the nucleus will evolve. If $r < r_C$, minimization of the free energy will cause the nucleus to shrink until it disappears. By contrast, if $r > r_C$, minimization of the free energy will cause the nucleus to grow until it takes over the entire system.

Problem: What is the critical radius r_C in a two-dimensional system?

Solution: For a 2D circular nucleus, there is a difference in the free energy density over the *area* of the nucleus, as well as a surface free energy associated with the *circumference* of the nucleus. Hence, the total free energy of the nucleus is $\Delta F = -\pi r^2 \Delta f + 2\pi r \sigma$. This expression has a maximum at the critical radius $r_C = \sigma/\Delta f$.

Problem: Is there a critical radius for a one-dimensional system?

Solution: A 1D nucleus is just a line segment of length $2r$ with two ends. There is a difference in the free energy density over the length of the nucleus, as well as surface free energy associated with the two ends. Hence, the total free energy is $\Delta F = -2r\Delta f + 2\sigma$. The important point to notice about this expression is that it does not have a maximum. Rather, it is always decreasing as a function of r. Physically, this result shows that a 1D nucleus of the stable phase can always reduce its free energy by growing; it can never reduce its free energy by shrinking. Hence, there is no critical radius in 1D.

In this scenario, the dynamic process of a first-order phase transition involves two steps. First, we have to wait for the formation of one nucleus, or several nuclei, larger than the critical radius. Second, we have to wait for the nucleus or nuclei to grow, so that the entire system will be filled with the new stable phase rather than the old metastable phase. This two-step process is called *nucleation and growth*.

The important point to remember about nucleation and growth is that it requires time. Hence, the behavior that you will observe in an experiment depends on the time scale of your experiment compared with the time scale required for nucleation and growth. If your experiment is done very slowly, then there will be enough time for the system to reach the stable state, through nucleation and growth, at every stage of the experiment. As a result, you will observe the stable equilibrium values of the order parameters. By comparison, if your experiment is done more quickly, then there will not be enough time for the system to reach the stable state, and you may observe the metastable values of the order parameters.

(You might ask: How much time is required? That is a difficult question, and there is no general answer. The answer varies tremendously depending on the specific system and on the temperature. As the temperature decreases, nucleation and growth requires much more time. We will discuss a specific case of that phenomenon in the next chapter.)

For our example of an Ising magnet, we are passing through the first-order phase transition by varying the applied magnetic field h. Hence, the experiment measures the magnetic order parameter M as a function of h at fixed temperature. If the field is changed sufficiently slowly, then the experiment will observe the equilibrium value of M, as calculated in Chap. 2. One example is shown in Fig. 7.5a. However, if the field is changed rapidly, then the magnet will not have enough time for nucleation and growth, and hence it will be stuck in the metastable state past the equilibrium phase transition. An example of this behavior is shown in Fig. 7.5b. As the field is increased from negative to positive values, M remains in the negative, metastable state for some time, until h reaches a substantial positive value. Conversely, as h is decreased from positive to negative values, M remains in the positive, metastable state until h reaches a substantial negative value. Hence, the plot shows a loop in M as a function of h. This phenomenon is called *hysteresis*, and the size of the hysteresis

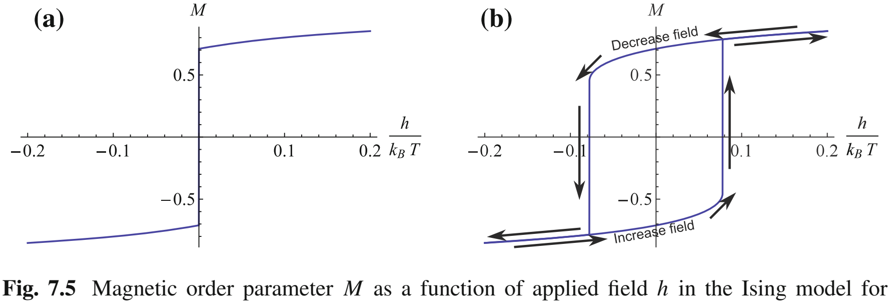


Fig. 7.5 Magnetic order parameter M as a function of applied field h in the Ising model for $T = 0.8T_C$. **a** Equilibrium order parameter with no hysteresis. **b** Hysteresis in the order parameter

loop indicates how much the system is out of equilibrium as it crosses the first-order transition.

As you might expect, the dynamic process of nucleation and growth has an analog in the liquid-gas system. The liquid-gas system has a first-order transition for temperatures and pressures below the critical point. We can cross this first-order transition by varying either temperature or pressure. Suppose we begin in the gas phase, and suddenly decrease the temperature or increase the pressure, so that the liquid phase has a lower free energy than the gas. After this change, the system can form a nucleus of the stable liquid phase surrounded by the metastable gas phase. Physically, the nucleus is just a tiny droplet of liquid. Equivalently, in the language of order parameter fields, we can say that the droplet is a region where the density field is near the stable liquid value, surrounded by a background where the density has the much lower gas value. If the radius of the droplet is greater than the critical radius, then the droplet tends to grow. Eventually, one or more droplets will grow until they take over the entire system.

The same process happens in reverse if we begin in the liquid phase, and suddenly increase the temperature or decrease the pressure, so that the gas phase has a lower free energy than the liquid. After the change, the system can form a nucleus of the stable gas phase surrounded by the metastable liquid. Physically, the nucleus is a tiny bubble of gas. If the radius of the bubble is greater than the critical radius, then the bubble tends to grow. Eventually, one or more bubbles will grow to fill the entire system.

7.2 Heterogeneous Nucleation

In the dynamic process described in the previous section, the system is homogeneous (or uniform), and hence a nucleus forms at random anywhere in the system. For this reason, the process may be called *homogeneous nucleation*. There is an important variation on this process, in which the system is not homogeneous; this variation is called *heterogeneous nucleation*.

In heterogeneous nucleation, the basic concept is that the system has some special positions that are favorable for formation of the new stable phase. In many cases, these special positions are impurities, such as dust particles. When a fluid is cooled from the gas phase to the liquid phase, the molecules may preferentially stick to the dust, and hence the a small liquid droplet will form around the particle. If the droplet is larger than the critical radius, then more molecules will stick to it, and hence it will grow. In this way, the dust particle can make it easier to form a critical nucleus, without the need to get over the high free energy barrier for homogeneous nucleation.

Heterogeneous nucleation is actually more common than homogeneous nucleation, presumably because most samples have many impurities like dust particles. Moreover, there can be other types of heterogeneity. The surface of a sample can serve as a favorable location for nucleation, especially if it is rough.

People sometimes intentionally put heterogeneities into a system in order to speed up first-order phase transitions. One common example is cloud seeding: If there is a drought on land, but there is water vapor in the clouds, people drop particles out of airplanes into the clouds. These particles serve as nucleation sites for the formation of droplets of liquid water. When the droplets grow large enough, they fall from the clouds as rain. Another example is the use of boiling chips: People may put small chips of a rough material into a liquid that is being heated. The surface of these chips serves as nucleation sites for the formation of bubbles of the gas phase. In this way, the chips make sure that the boiling transition will occur at the equilibrium transition temperature. They prevent the formation of a metastable liquid above the equilibrium transition temperature, which might suddenly form a large bubble, creating a hazardous condition.

Incidentally, the science of nucleation and growth is an important theme in the famous American novel *Cat's Cradle,* by Kurt Vonnegut, from 1963. The author's older brother Bernard Vonnegut was a chemist at General Electric, who discovered that silver iodide is especially effective for cloud seeding. Kurt used this concept for the rather disturbing fictional story of "ice-nine" nucleation in the novel.

Further Reading

The study of nucleation and growth is an essential part of materials science, and it is discussed in many textbooks. Some examples are:

1. K.F. Kelton, A.L. Greer, *Nucleation in Condensed Matter: Applications in Materials and Biology* (Pergamon, 2010)
2. D. Kashchiev, *Nucleation: Basic Theory with Applications* (Butterworth-Heinemann, 2000)
3. V.I. Kalikmanov, *Nucleation Theory* (Springer, 2013)

Chapter 8
Solids: Crystals and Glasses

Abstract This chapter discusses the physics of solids, as distinct from gases and liquids. It begins with an introduction to crystals, emphasizing the positional and orientational order of the crystalline phase. As a special case, it considers the efficient packing of spherical particles in close-packed crystals, leading to the crystallization of hard spheres. It then goes on to discuss elasticity and viscosity, identifying the mechanical properties that distinguish all solids from liquids. Based on these mechanical properties, it shows how glasses can form as non-equilibrium, non-crystalline solids.

So far in this book, we have discussed all phenomena in terms of two physical examples: the Ising magnet and the liquid-gas system. It is the time to move on to another physical example: solids!

8.1 What is a Solid?

We all have some intuitive feeling about what is a solid. If we had to explain this impression to a child, we might put it in one of the two different ways:

1. If the child is very young, we might say: A solid is a material that you can hold in your hand; it does not flow out between your fingers. Later in this chapter, we will see that a scientific way to express this concept is: A solid is a material that has a nonzero *shear modulus*.
2. If the child is somewhat older, we might say: A crystalline solid has a shape with sharp facets. Or even: A crystalline solid is a material that diffracts X-rays, leading to sharp Bragg spots.

These two statements are quite different from each other, and they refer to different classes of materials. Statement #1 is very broad, and we can take it as a general definition of what we mean by solids. Statement #2 is much more specific, and it describes a particular subset of solids, known as *crystals*. Not all solids are crystals.

J.V. Selinger, *Introduction to the Theory of Soft Matter*,
Soft and Biological Matter, DOI 10.1007/978-3-319-21054-4_8

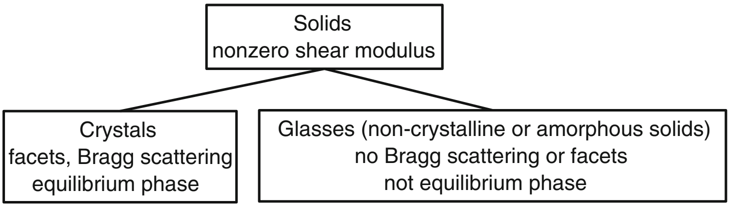


Fig. 8.1 Schematic diagram of what is meant by solids

The materials that satisfy statement #1 but not statement #2 are known as *non-crystalline solids,* or *amorphous solids,* or *glasses.* Figure 8.1 shows a schematic diagram of this distinction.

In this chapter, we will first consider crystals, because they are an equilibrium phase like gases and liquids. We will then discuss the concepts of elasticity and viscosity, which are needed in order to make the distinction between solids and liquids based on the response to a shear. In the end, we will use these concepts to describe glasses.

8.2 Crystals

A crystal is a phase in which atoms are arranged in a periodic lattice. The lattice points form a regular array in 2D or 3D, as shown in Figs. 8.2 and 8.3. The atoms do not need to be exactly located on the lattice points. Indeed, at any nonzero temperature, the atoms are constantly vibrating, so they are never *exactly* on the lattice points. However, the atoms are vibrating about the lattice points, so each atom remains close to its lattice point. This means that there is a great regularity in the positions of the atoms. The repeating length scale (known as the *lattice constant*) of the crystal

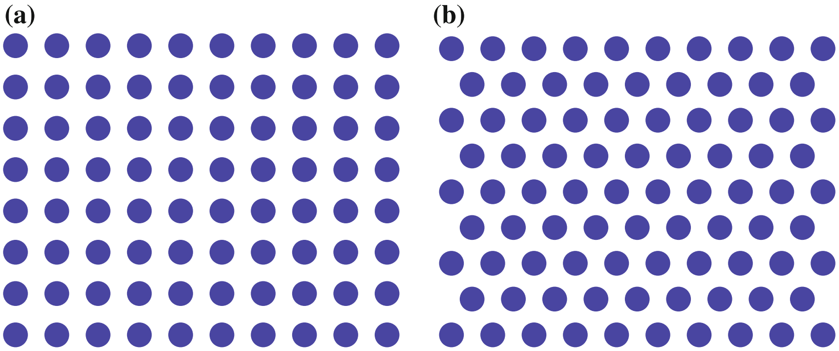


Fig. 8.2 Examples of common 2D crystal structures. **a** Square. **b** Hexagonal (or triangular)

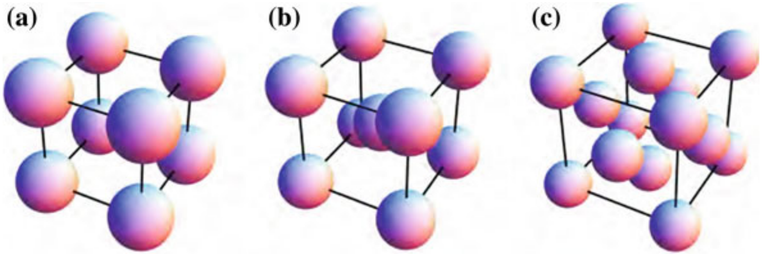


Fig. 8.3 Examples of common 3D crystal structures. **a** Simple cubic (sc). **b** Body-centered cubic (bcc). **c** Face-centered cubic (fcc) (Interactive version at http://www.springer.com/cda/content/document/cda_downloaddocument/Selinger+Interactive+Figures.zip?SGWID=0-0-45-1509169-p177545420.)

keeps repeating over great distances. Suppose we have a crystal with a lattice constant of 1 nm, and one of the crystal axes is in the east–west direction. If we begin at some test atom and move to the east, we find another atom at 1 nm, another at 2 nm, ..., another at 100 nm, another at 101 nm, ...

Crystallography is a very well-developed area of mathematics and physics, and it is well described in books on solid-state physics. I will not attempt to summarize all the principles of crystallography here, but I will emphasize two of the main conclusions. First, all of the possible crystal structures in 2D and 3D have been classified. The structures shown in Figs. 8.2 and 8.3 are a few of the most common structures, but there are many more. Second, the periodic repeating lattice is what gives the remarkable features of crystals mentioned in the introduction—facets and sharp Bragg peaks in the X-ray diffraction. These features are macroscopic manifestations of the periodic crystal structure.

What can we say about crystals from the perspective of order, disorder, and statistical mechanics? To begin, let us compare crystals with liquids (and equivalently gases) in terms of *symmetry*. We can consider both positions and orientations within the material:

Positions

Crystals are highly symmetric, in the sense that all lattice sites are equivalent to each other.[1] We are equally likely to find an atom near any of the lattice sites. Mathematically, we can say that crystals are symmetric under *translations* by any lattice vector, which shifts lattice sites onto other lattice sites.

By comparison, in liquids, *all* positions are equivalent to each other. We are equally likely to find an atom at any position, not just near certain special positions. Hence, liquids are symmetric under translations by any vector at all. Although crystals are highly symmetric, liquids are even more symmetric!

[1] Here, we neglect the effects of surfaces and assume that the crystal goes on to infinity in all directions.

Orientations

Crystals have several special orientations called *crystalline axes*, which are the directions along which we can find rows of atoms. Some of these crystalline axes are equivalent to each other. For example, in a simple cubic lattice, the x, y, and z directions are all equivalent to each other. Hence, the simple cubic lattice is symmetric under certain *rotations* by 90°, which take x into y, y into z, or z into x.

By comparison, in liquids, *all* orientations are equivalent to each other. We are equally likely to find atoms along any direction in the liquid, not just along certain special directions. Mathematically, we can say that liquids are symmetric under any rotation. Again, although crystals are highly symmetric, liquids are even more symmetric.

This symmetry comparison is sometimes counter-intuitive for students, who are accustomed to thinking of crystals as very symmetric objects. Indeed, sometimes students ask: What happens if we look at a snapshot of a liquid, showing the positions of the atoms at some particular time. In this snapshot, the atoms are at random positions, and we do not see any symmetry at all. The response is: We are talking about a statistical symmetry, averaged over many configurations of the liquid, averaged over time. On this statistical basis, all positions within the liquid are equivalent to each other, and all directions are equivalent to each other.

We can see this symmetry comparison, for example, in the optical properties of crystals compared with liquid and gases. If we shine light through a quartz crystal, the optical properties are the same for light shining along certain special axes; the optical properties are different for light along other axes. By comparison, if we shine light through air, the optical properties are the same for light in any direction whatsoever. In this sense, air is more symmetric than a quartz crystal.

Next, let us compare crystals with liquids (and equivalently gases) in terms of *order*. Recall the discussion of order and symmetry in the Ising model from Sect. 2.3. At this point, we saw that magnetic order is the *opposite* of symmetry; order means that the magnet has selected a special direction, up or down, for the net magnetization. The same point is true for the crystal. Because the crystal has less positional symmetry than the liquid, we must say that the crystal has more positional order. Of all the equivalent positions in the liquid, the crystal has selected certain special points for the crystalline lattice. (For terminology, we would say that the liquid is *uniform*, while the crystal is *nonuniform* and *periodic*.) Likewise, because the crystal has less orientational symmetry than the liquid, the crystal has more orientational order. Of all the equivalent directions in the liquid, the crystal has selected certain special directions for the crystalline axes. (For terminology, the liquid is *isotropic,* while the crystal is *anisotropic*.) Hence, the transition from liquid to crystal involves *breaking symmetry* and *acquiring order*—both positional order and orientational order.

Because the system acquires order when it freezes from liquid to crystal, we would like to describe the order by an order parameter. To construct an order parameter, consider the density as a function of position. Figure 8.4 shows a sample plot of the density $\rho(\boldsymbol{r})$ for a 2D square lattice. We can see that the density is peaked at the lattice sites, and it drops to minima between the lattice sites, so that the density plot

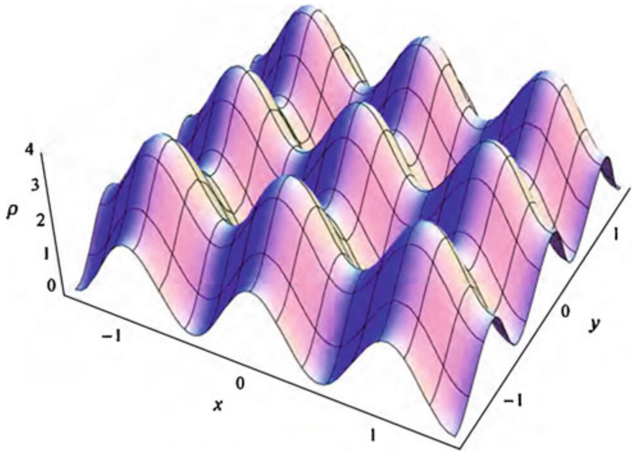


Fig. 8.4 Density in a 2D square lattice as a function of position (Interactive version at http://www.springer.com/cda/content/document/cda_downloaddocument/Selinger+Interactive+Figures.zip?SGWID=0-0-45-1509169-p177545420.)

has the shape of an egg carton. Because it is a periodic function, it can be written as a Fourier series

$$\rho_{\text{crystal}}(\boldsymbol{r}) = \rho_0 + \sum_{\boldsymbol{q}} \rho_{\boldsymbol{q}} e^{i\boldsymbol{q}\cdot\boldsymbol{r}}. \tag{8.1}$$

In this sum, the wavevectors are special wavevectors associated with the particular lattice.[2] In the particular example of Fig. 8.4, they are the wavevectors $(2\pi/a)\hat{\boldsymbol{x}}$ and $(2\pi/a)\hat{\boldsymbol{y}}$, where a is the square lattice spacing. By comparison, in the liquid phase, the density is uniform and can be written as

$$\rho_{\text{liquid}}(\boldsymbol{r}) = \rho_0. \tag{8.2}$$

Hence, the distinction between crystal and liquid can be described by the set of Fourier coefficients $\rho_{\boldsymbol{q}}$. These coefficients are zero in the liquid phase, and at least some of them become nonzero in the crystal phase. In general, these coefficients are complex numbers. The magnitude of the complex numbers represents the amplitude of the density wave; it shows whether there is a small or large difference between the densities at the maxima and the minima. The phase of the complex numbers represents the positions of the density maxima; the spontaneous symmetry breaking in the choice of where to put the maxima corresponds to a random choice of the phases of $\rho_{\boldsymbol{q}}$. Thus, these density order parameters describe both the magnitude and direction of the positional symmetry breaking. In this respect, they are a generalization of the Ising order parameter M.

[2] In the general formalism of solid-state physics, the wavevectors in this sum are the *reciprocal lattice vectors* of the particular lattice; I will not discuss that concept further in this book.

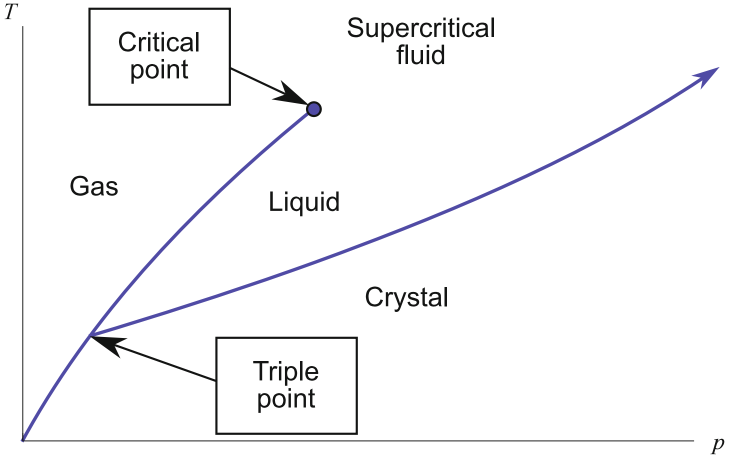


Fig. 8.5 Generic phase diagram for the crystal, liquid, and gas phases, in terms of pressure and temperature

If a crystal has positional order defined by the coefficients $\rho_{\boldsymbol{q}}$, then the wavevectors $\boldsymbol{q}$ already contain information about the orientations of the crystalline axes. For this reason, there is no need for a separate orientational order parameter to describe the crystal. (An orientational order parameter will come later, in Chap. 10.)

At this point, I would like to present a simple theory of the freezing transition from liquid to crystal, analogous to the van der Waals theory for the gas-liquid transition. Unfortunately, no such simple theory exists. The freezing transition is more complicated, and it varies from material to material. In particular, it depends on the type of interaction between atoms—different types of crystals arise from directional covalent bonds compared with ionic bonds or other isotropic interactions between atoms. Even so, we can at least consider the generic phase diagram for crystals along with liquids and gases.

Figure 8.5 shows the generic phase diagram in terms of pressure and temperature. The crystal phase normally occurs at a lower temperature than the liquid phase because the crystal has more order. At high temperature, entropy favors the more disordered liquid phase; at low temperature, energy favors the more ordered crystal phase. Likewise, the crystal normally occurs at a higher pressure than the liquid because the crystal has a higher density (lower volume per particle) in most materials.[3] At low temperatures and pressures, the gas, liquid, and crystal phases come together at the *triple point*. The triple point must not be confused with the critical point where the distinction between liquid and gas vanishes; it is a completely different type of point. Below the triple point, there is no liquid phase; rather, there is a direct transition between crystal and gas.

[3] Water is a notable exception to this general rule, because crystalline ice actually has a lower density than liquid water.

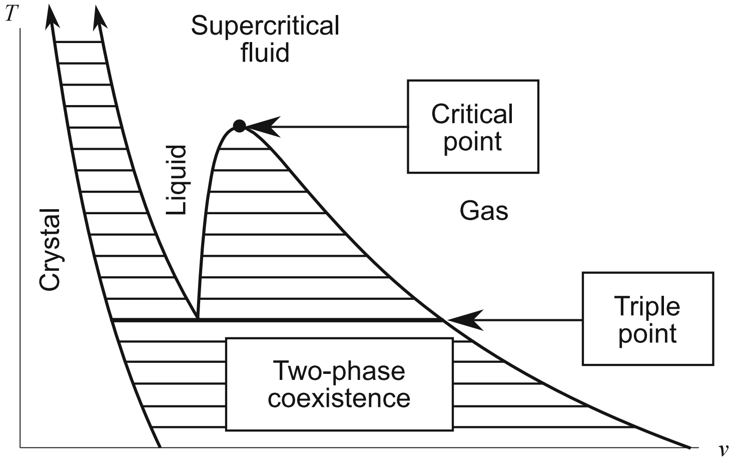


Fig. 8.6 Generic phase diagram for the crystal, liquid, and gas phases, in terms of volume per particle and temperature

At high temperatures and pressures, the transition between crystal and liquid goes on forever; it does not terminate in a critical point. This absence of a critical point is an important consequence of symmetry: The distinction between crystal and liquid is a fundamental symmetry difference—the liquid has translational and rotational symmetries that the crystal lacks. This symmetry is either present or absent. For this reason, there must always be a phase transition between crystal and liquid, where the symmetry is broken; this phase transition cannot terminate in a critical point. By contrast, liquid and gas are the same type of phase, with the same symmetry; the distinction between them is a quantitative difference in the density. Because this difference is only quantitative, it can vanish at the critical point.[4]

Figure 8.6 shows the generic phase diagram in terms of volume per particle and temperature. In this version of the phase diagram, we can see that there are regions of two-phase coexistence between gas and liquid, liquid and crystal, and gas and crystal. The coexistence region between gas and liquid terminates at the critical point, but the coexistence region between liquid and crystal goes on forever. Furthermore, in this version of the phase diagram, the triple point becomes a line of three-phase coexistence, at a specific triple-point temperature, with three different volumes per particle for gas, liquid, and crystal.

[4] For more advanced readers: The only way to create a critical point in the crystal-liquid transition would be to apply a symmetry-breaking field that directly induces positional order. For example, suppose we apply a laser interference pattern, with the same spatial periodicity as the crystal. Under this symmetry-breaking field, the liquid phase would have a slight degree of positional order, and the crystal phase would have enhanced positional order. If the field strength is sufficiently great, the difference in positional order could vanish at a critical point. This is a fairly exotic situation, which does not occur in the ordinary study of phase transitions.

8.3 Close-Packed Crystals, Crystallization of Hard Spheres

One specific type of crystal structure deserves special attention. In many systems, the crystal structure that minimizes the free energy is whatever crystal structure fills space most efficiently. By filling space efficiently, we just mean putting the most particles into the smallest possible volume, with the least wasted space between the particles. The most efficient packing is called a *close-packed* crystal. Such crystals typically occur if the interaction between particles is isotropic, so that the particles are effectively spheres in 3D, or disks in 2D.

To characterize how efficiently a crystal structure fills 3D space, people define the *packing fraction* or *volume fraction* ϕ as the fraction of volume that is actually taken up by spherical particles, compared with the total volume in the lattice:

$$\phi = \frac{V_{\text{particles}}}{V_{\text{total}}}. \tag{8.3}$$

For a 2D crystal, the corresponding *area fraction* is the ratio of areas instead of volumes.

We can calculate some examples of packing fractions for inefficient crystal structures compared with close-packed crystals in 2D and 3D:

In 2D, one simple lattice is the square lattice. To calculate the packing fraction for the square lattice, we shrink the lattice constant until neighboring disks are just touching each other, without overlapping, as shown in Fig. 8.7a. In this case, the lattice constant a equals the particle diameter. The repeating unit cell of this lattice is the black square in the figure. Because the entire crystal consists of copies of the unit cell, it is sufficient to calculate the numerator and denominator of Eq. (8.3) for a single unit cell. For the numerator, there is one particle per unit cell (four quarter-particles), and the particle area is $\pi(a/2)^2$. For the denominator, the total area of the unit cell is a^2. Hence, the packing fraction for the square lattice is $\phi = \pi/4 \approx 0.785$.

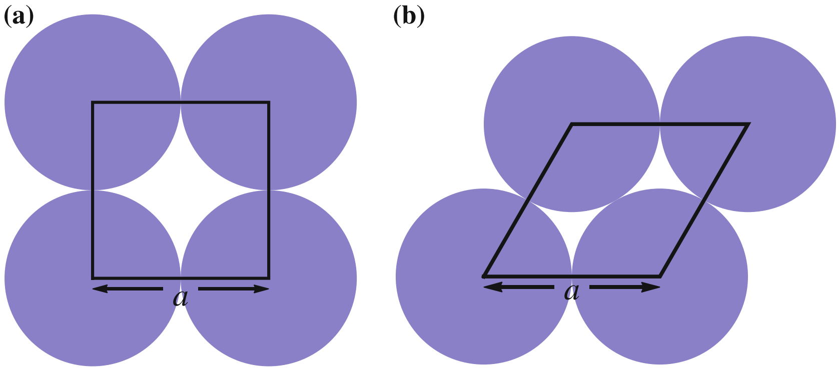


Fig. 8.7 Calculation of the packing fraction for 2D lattices. **a** Square. **b** Hexagonal (or triangular)

The packing fraction is a pure number between 0 and 1, which does not depend on the lattice constant a, but only on the lattice type. This number implies that 78.5 % of the area is used for particles, and 21.5 % of the area is wasted for the empty space between the particles.

By comparison, the most efficient packing in 2D is the hexagonal lattice. This is the crystal structure that you would obtain, for example, if you pack identical coins together on a table. To calculate the packing fraction for the hexagonal lattice, we again shrink the lattice constant until neighboring particles are just touching, so that the lattice constant a equals the particle diameter. The unit cell of this lattice is the parallelogram shown in Fig. 8.7b. For the numerator of Eq. (8.3), there is one particle per unit cell, and the particle area is $\pi(a/2)^2$. For the denominator, the area of the parallelogram is $a^2\sqrt{3}/2$. Hence, the packing fraction for the 2D hexagonal lattice is $\phi = \pi/(2\sqrt{3}) \approx 0.907$. Thus, the hexagonal lattice is a substantially more efficient packing than the square lattice, with less wasted space. Indeed, this difference is clearly visible as less empty space in Fig. 8.7b than in 8.7a.

In 3D, we can first consider the simple cubic (sc) lattice. When we shrink the lattice constant until neighboring spheres are just touching, we obtain the structure shown in Fig. 8.8a, with the lattice constant a equal to the sphere diameter. The unit cell is the black cube, and it contains one sphere (eight eighth-spheres). The volume used by the sphere is $(4\pi/3)(a/2)^3$, and the total volume of the unit cell is a^3. Hence, the packing fraction of the simple cubic lattice is $\phi = \pi/6 \approx 0.524$.

By comparison, the face-centered cubic (fcc) lattice is a close-packed crystal structure in 3D. Figure 8.8b shows this structure, when the lattice constant is shrunk until neighboring spheres are just touching. In this case, the cubic lattice constant a is not equal to the sphere diameter d. Rather, by considering the diagonal across a face, we can see that $a = d\sqrt{2}$. The cubic unit cell contains four spheres (eight eighth-spheres at the corners and six half-spheres on the faces), and each sphere has

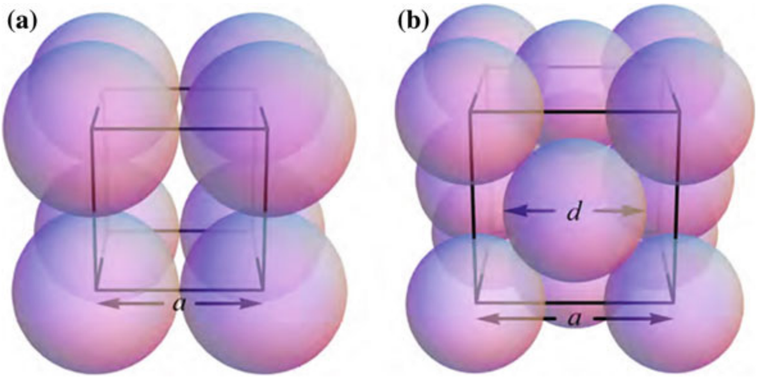


Fig. 8.8 Calculation of the packing fraction for 3D lattices. **a** Simple cubic (sc). **b** Face-centered cubic (fcc) (Interactive version at http://www.springer.com/cda/content/document/cda_downloaddocument/Selinger+Interactive+Figures.zip?SGWID=0-0-45-1509169-p177545420.)

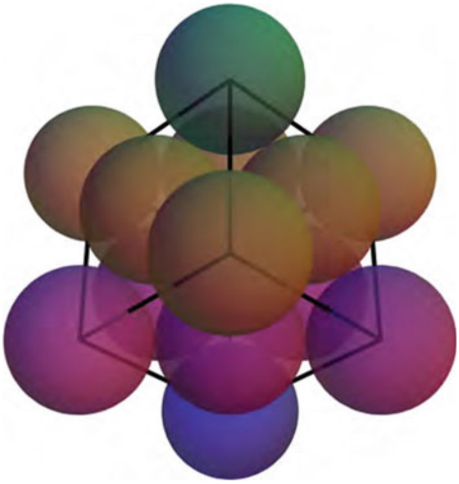


Fig. 8.9 Alternative view of the fcc lattice, showing the hexagonal layers (Interactive version at http://www.springer.com/cda/content/document/cda_downloaddocument/Selinger+Interactive+Figures.zip?SGWID=0-0-45-1509169-p177545420.)

a volume of $(4\pi/3)(d/2)^3$. The total volume of the unit cell is a^3. Hence, the packing fraction of the fcc lattice is $\phi = \pi/(3\sqrt{2}) \approx 0.740$, much higher than the simple cubic.

To understand why the fcc lattice is so efficient, consider the alternative visualization in Fig. 8.9. This visualization shows exactly the same spheres as Fig. 8.8b, but the cube has been tilted so that the body diagonal is vertical, and the spheres have been colored by layer. Here, we can see that each layer of the fcc lattice is a 2D hexagonal lattice of spheres, which is the most efficient packing in 2D. Each layer of spheres is placed over the depressions in the layer below. This is exactly the way that you would stack oranges in a supermarket! You should rotate the interactive version of the figure in 3D in order to see this arrangement more clearly.

There are three possible ways to place the hexagonal layers, which are conventionally called A, B, and C. In the fcc lattice, the layers follow the repeating pattern ABCABC…, so the the fourth layer is directly above the first layer. The hexagonal layers might be arranged in different patterns, leading to different lattices. In particular, the hexagonal close-packed (hcp) lattice has the layers in the pattern ABABAB…. Any such pattern of hexagonal layers has the same packing fraction $\phi = \pi/(3\sqrt{2}) \approx 0.740$. This has been proven to be the most efficient packing of spheres in 3D.

We noted earlier that close-packed lattices, such as fcc, generally occur when the interaction between particles is isotropic. One interesting special case of isotropic particles is *hard spheres*. Hard spheres are particles that have no interaction other than excluded-volume repulsion, which prevents them from overlapping. In other words, the interaction potential between particles i and j is

$$V_{ij} = \begin{cases} 0 & \text{if no overlap,} \\ \infty & \text{if overlap.} \end{cases} \tag{8.4}$$

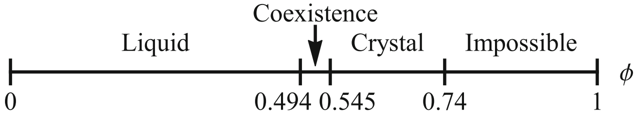


Fig. 8.10 Phase diagram for hard spheres in 3D, as a function of the single variable ϕ

The hard-sphere system is peculiar because it does not have any interaction energy—the potential energy is zero for any allowed configuration of the spheres (with no overlap), and hence the expectation value of the potential energy is also zero. As a result, the hard-sphere system is entirely controlled by entropy. One might expect that a system with only entropy and no energy could never crystallize, because entropy would always favor disorder. Surprisingly, it actually does crystallize!

For hard spheres, the phase diagram depends on only one variable, the volume of the container *per* sphere in comparison with the volume *of* each sphere, or equivalently the packing fraction ϕ. Figure 8.10 shows the phase diagram as a function of this single variable. For $0 < \phi < 0.494$, the system is in the liquid phase. For $0.494 < \phi < 0.545$, it shows two-phase coexistence between liquid and fcc crystal, with the fractions of each phase given by the lever rule. For $0.545 < \phi < 0.740$, it is the fcc crystal phase. It is impossible to have $\phi > 0.740$, because this is the close-packed packing fraction for spheres in 3D.

We might well ask *why* the hard-sphere system crystallizes, considering that it is controlled entirely by entropy. The answer is that there is a balance between different types of entropy. Clearly, the liquid has long-range entropy, because it has the highest possible symmetry, and does not have any special positions or orientations. However, when the density of spheres becomes high, the liquid has a problem: the spheres are jammed together in a disordered way, and they do not have room for small displacements around their positions. If the system crystallizes into the very efficient fcc lattice, then spheres are not jammed against their neighbors, and they have room for small displacements around their lattice sites. These small displacements are a type of short-range entropy, and they favor the crystal phase. The competition between long-range and short-range entropy leads to crystallization as a function of density. In Chap. 10, we will see that a similar type of transition can occur with the orientations of hard rods.

8.4 Elasticity and Viscosity

So far in this chapter, we have discussed crystals, which are one important type of solids. In order to discuss solids in general, we must consider their elastic response to forces, and see how this elastic response differs from liquids.

The simplest type of mechanical response, which is certainly familiar to physics students, is Hooke's law for springs. Consider the spring illustrated in Fig. 8.11a. We apply a force F on one end, and a compensating force on the other end so that the spring will not fly away. (If the spring is mounted on a wall, then we apply a force F

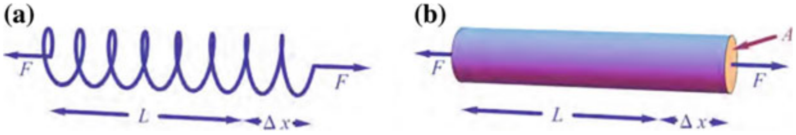


Fig. 8.11 **a** Stretching of a spring by a force F. **b** Stretching of an elastic rod with initial length L and cross-sectional area A (Interactive version at http://www.springer.com/cda/content/document/cda_downloaddocument/Selinger+Interactive+Figures.zip?SGWID=0-0-45-1509169-p177545420.)

on one end, and the wall applies a compensating force on the other end.) In response to this force, the spring extends from its natural length L by a distance Δx. Hooke's law states that the force is related to the extension by

$$F = k\Delta x, \tag{8.5}$$

with some spring constant k. The same relationship is true for any solid rod, as shown in Fig. 8.11b. If we apply a force to both ends, then the rod extends following Hooke's law, with some spring constant.

The important point to emphasize about Hooke's law is that the spring constant depends on the size and shape of the rod, as well as on the type of material that the rod is made of. If we double the cross-sectional area A of the rod, we get twice the force for the same Δx (just as if we put two springs in *parallel*). Hence, we see that the spring constant k is proportional to A. Likewise, if we double the natural length L of the rod, we get twice the extension Δx for the same force (just as if we put two springs in *series*). Hence, we see that k is inversely proportional to L.

It is often useful to separate the effects of the rod dimensions A and L from the effects of the material. For this reason, we can write the spring constant as

$$k = \frac{EA}{L}, \tag{8.6}$$

where the coefficient E is called the *Young's modulus*. The Young's modulus does not depend on the the size and shape of the rod; rather, it is a characteristic of the material, which describe how easy or difficult it is to stretch the material. It is measured in Pascals (Pa), and it ranges from 0.01 to 0.1 GPa for rubber, to 70 GPa for aluminum, to 1000 GPa for diamond.

In terms of Young's modulus, we can rewrite Hooke's law as

$$F = \frac{EA}{L}\Delta x. \tag{8.7}$$

We can then rearrange terms to obtain

$$\frac{F}{A} = E\frac{\Delta x}{L}. \tag{8.8}$$

This form of Hooke's law suggests that it is useful to normalize the force F by the cross-sectional area A, and normalize the extension Δx by the natural length L. Indeed, these are very standard concepts in elasticity. The normalized force is called the *stress* σ,

$$\sigma = \frac{F}{A}, \tag{8.9}$$

and the normalized extension is called the *strain* e,

$$e = \frac{\Delta x}{L}. \tag{8.10}$$

We can see that σ has units of Pascals, while e is dimensionless. In terms of these variables, Hooke's law takes the form

$$\sigma = Ee. \tag{8.11}$$

This equation emphasizes the properties of the material, while putting the size and shape of the rod into the definitions of stress and strain.

A linear extension is not the only way to deform a material. Another important type of deformation is a *shear,* as illustrated in Fig. 8.12. In a shear, we apply a horizontal force to the top of the material, and an opposite horizontal force to the bottom. As a result, the material deforms by a distance Δx, from an initially rectangular shape into a parallelepiped. For this problem, the shear strain is defined as $e = \Delta x/L$, and the shear stress is defined as $\sigma = F/A$. Be careful to notice which directions have the length L and area A: L is the length of the material from bottom to top, across which the force and extension are varying. A is the area of the top and bottom, over which the force is applied, perpendicular to L.

For a solid, the shear stress is linearly proportional to the shear strain:

$$\sigma = Ge, \tag{8.12}$$

where G is the *shear modulus* of the material. (It is sometimes denoted by the letter μ.) It is a different material parameter than Young's modulus, because it describes the response of a material to a different type of deformation. It is also measured in Pascals, and it is normally smaller than Young's modulus. Typical values are 0.001 GPa for rubber, 30 GPa for aluminum, and 500 GPa for diamond.

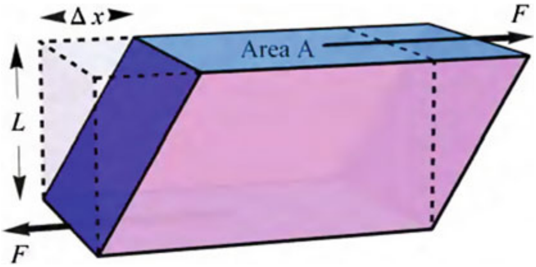


Fig. 8.12 Shear deformation of a material (Interactive version at http://www.springer.com/cda/content/document/cda_downloaddocument/Selinger+Interactive+Figures.zip?SGWID=0-0-45-1509169-p177545420.)

Now we can see the crucial difference between solids and liquids: When we apply a shear stress to a solid, the top surface shifts by Δx and then it stops. By contrast, when we apply a shear stress to a liquid, the top surface just keeps flowing without limit. A liquid does not have any steady-state shear strain; instead, it has a steady-state shear strain *rate*. The shear strain rate is the velocity of the top surface, normalized by the length L, or equivalently it is the derivative of the strain with respect to time:

$$\dot{e} \equiv \frac{de}{dt} = \frac{d(x/L)}{dt} = \frac{1}{L}\frac{dx}{dt} = \frac{v}{L}. \tag{8.13}$$

In this expression, $\dot{e}$ (pronounced "e-dot") is a traditional notation for the derivative of e with respect to time. The shear strain rate is linearly proportional to the shear stress:

$$\sigma = \eta\dot{e}, \tag{8.14}$$

where η is the *viscosity* of the liquid. The SI unit for viscosity is Pascal seconds (Pa s), and it is also commonly reported in the non-SI unit Poise, where 1 Poise is 0.1 Pa s. As examples, water has a viscosity of 0.001 Pa s at room temperature, while the viscosity of honey is about 2 Pa s.

To distinguish between solids and liquids, we can say: A solid has a nonzero shear modulus, but a liquid has no shear modulus (or equivalently, the shear modulus of a liquid is zero). This is exactly the distinction that people have in mind when they say: You can hold a solid in your hand, but a liquid flows out between your fingers. As an example, Fig. 8.13 shows a schematic illustration of someone holding a material on top of his fingers. We can see that it looks like two copies of the shear experiment

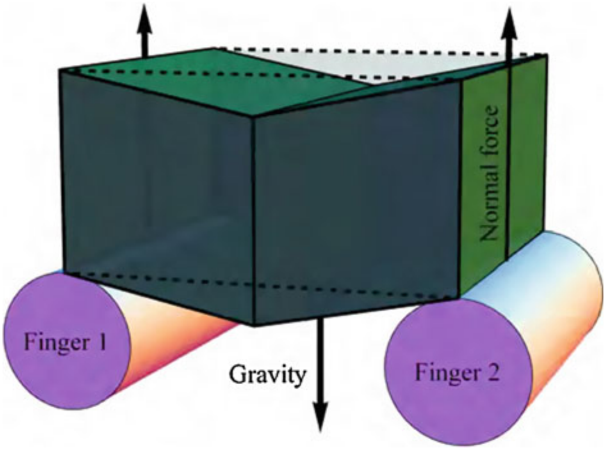


Fig. 8.13 Schematic illustration of someone holding a material on top of his fingers to determine whether it is a solid or a liquid (Interactive version at http://www.springer.com/cda/content/document/cda_downloaddocument/Selinger+Interactive+Figures.zip?SGWID=0-0-45-1509169-p177545420.)

of Fig. 8.12, turned on its side. Gravity provides a force pulling down on the center of the material, and the fingers provide a normal force pushing up on the two sides, leading to a shear stress σ. If the material is a solid, it will deform until it reaches a shear strain of $e = \sigma/G$, and then it will stop. By contrast, if the material is a liquid, it will flow downward between the fingers with a shear rate of $\dot{e} = \sigma/\eta$, and it will not stop until all the material has flowed out onto the floor.

This is also the distinction that people have in mind when they say: A liquid conforms to the shape of its container, but a solid keeps its own shape. If we put a sample into a container, the walls exert many shear stresses on the sample. If the sample is a solid, it will only deform by some small shear strains, and hence it will (approximately) keep its own shape. By contrast, if the sample is a liquid, it will flow until it reaches the shape of its container, at which point the shear stresses will vanish.

I should mention that the theory in this section describes idealized solids and liquids. The ideal solid is called a *Hookean solid* (after Robert Hooke), and the ideal liquid is called a *Newtonian liquid* (after Isaac Newton).[5] Real materials can be more complex than these idealizations in at least two ways:

First, the idealizations describe the steady-state shear of a solid or the steady-state flow of a liquid. Materials can have interesting time-dependent properties before they reach the steady state. For example, a material might respond elastically (like a solid) on short time scales, but flow viscously (like a liquid) on long time scales. If a material displays both elastic and viscous behaviors on different time scales, it is called *viscoelastic*. In the theory of viscoelasticity, the shear modulus G and viscosity η can be combined into a single frequency-dependent complex modulus $G(\omega)$. (For readers who have studied AC electric circuits, this complex modulus is quite analogous to the frequency-dependent complex impedance $Z(\omega)$, which combines the resistance, capacitance, and inductance.)

Second, the idealizations assume that the shear stress is linearly proportional to the shear strain (for a solid) or shear strain rate (for a liquid). This linear relationship is a good approximation for *small* stresses and strains. However, when stresses and strains become large, their relationship can be nonlinear. For this reason, a tremendous amount of effort goes into measuring the stress–strain curves for real materials.

I should also mention that the theory of elasticity is a very interesting and complex subject. In general, the stress and strain can both be represented by tensors, which show all the possible orientations for force gradients and displacement gradients within a material. The mathematical formalism of elasticity is similar to electrodynamics, except that elasticity uses tensors while electrodynamics uses vectors. That subject is beyond the scope of this book, but you should be aware that it exists.

[5] As a historical note, Robert Hooke (1635–1703) and Isaac Newton (1642–1726) were both great scientists, but they were bitter enemies. According to legend, after Hooke died, Newton had all portraits of him destroyed, so we do not know what Hooke looked like.

8.5 Microscopic Interpretation of Viscosity

In the previous section, we presented a description of elasticity and viscosity, but not an explanation for these behaviors. In this section, we will present a simple microscopic model for viscous flow. It will give a prediction for how the viscosity depends on temperature.

In this model, the basic concept is that each atom is trapped in a cage of its neighbors. As an example, Fig. 8.14 shows a cluster of atoms within a fluid. The test atom at the center, shown in blue, is surrounded by neighboring atoms, shown in yellow. All of these atoms are constantly moving back and forth. As long as the blue atom is trapped inside the cage of yellow atoms, it vibrates within the cage, and the cage tends to push it back to the center. However, if the blue atom ever escapes from the cage, it forms a new cluster with the atoms at its new position, and it no longer has any tendency to return to its original position.

Now suppose we apply a shear stress σ to the material. As long as the atoms are trapped within their cages, the shear stress just deforms each of the cages. This deformation induces a displacement of the atoms—a specific finite displacement—and then it stops, because of the energy cost of pushing the atoms away from their equilibrium positions. As a result, the material has a solid-like elastic response to the shear stress, with a shear strain of $e = \sigma/G_0$, where G_0 is the instantaneous shear modulus on this short time scale. If we remove the shear stress, each atom tends to return to its original position, at the center of its own cage, and hence the shear strain goes back to 0.

By contrast, if we wait long enough for the atoms to escape from their cages, the behavior is quite different. Suppose τ is the relaxation time, i.e., the typical time required for an atom to escape from its cage. If we apply the shear stress for a time τ, each atom moves out of its cage to a new position, and it forms a new cage. After another time τ, each atom moves to yet another position. Hence, the material

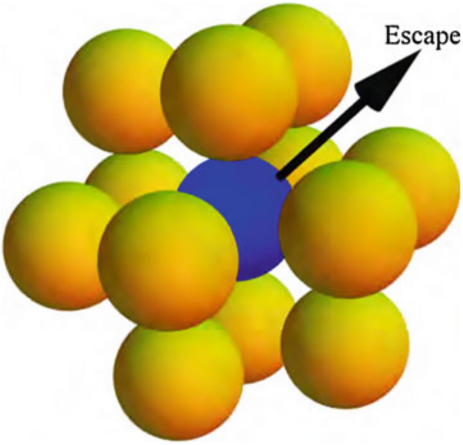


Fig. 8.14 Escape of a test atom (shown in *blue*) from a cage of its neighbors (shown in *yellow*), as part of a microscopic model for viscosity (Interactive version at http://www.springer.com/cda/content/document/cda_downloaddocument/Selinger+Interactive+Figures.zip?SGWID=0-0-45-1509169-p177545420.)

continues flowing for as long as the stress is applied, leading to a liquid-like viscous response. If we remove the shear stress, there is no tendency for the atoms to return to their original positions, and hence the strain does not go back to 0.

As a rough estimate, we can say that the material experiences a shear strain of $e = \sigma/G_0$ in each time τ. As a result, the average shear strain rate is $\dot{e} = \sigma/(G_0\tau)$. Hence, the effective viscosity η is determined by the instantaneous shear modulus G_0 and the relaxation time τ:

$$\eta = G_0\tau. \tag{8.15}$$

This argument provides a way for us to think about different materials in terms of the relaxation time. In a perfect crystal, any rearrangement of the atoms requires large-scale motion of all the atoms in the lattice, and hence $\tau \to \infty$. The perfect crystal can never flow. In a crystal with defects, τ is large but finite. In a liquid, τ is small, and hence the liquid can flow in a very short time, compared with the time scale of an experiment.

Now let us consider how τ depends on temperature. Because τ is the typical time for an atom to escape from its cage, τ^{-1} is the typical escape rate, i.e., escapes per unit time. To understand this escape rate, we can compare it with prisoners trying to escape from jail. The escape rate τ^{-1} (prisoners escaping per unit time) is equal to the attempt frequency ν (how often prisoners *try* to escape) times the probability of success:

$$\tau^{-1} = \nu p_{\text{success}}. \tag{8.16}$$

The same is true for atoms escaping from their cages. Here, ν is a vibration frequency that indicates how often an atom approaches the cage walls, and p_{success} is given by a Boltzmann factor for escape over a barrier,

$$p_{\text{success}} = e^{-F_{\text{barrier}}/k_B T}. \tag{8.17}$$

Combining these equations, our estimate for the relaxation time is

$$\tau = \nu^{-1} e^{F_{\text{barrier}}/k_B T}, \tag{8.18}$$

and our estimate for the viscosity is

$$\eta = \frac{G_0}{\nu} e^{F_{\text{barrier}}/k_B T}. \tag{8.19}$$

Equation (8.19) gives a prediction for how the viscosity depends on temperature: through an exponential factor. To be sure, the three parameters G_0, ν, and F_{barrier} may also depend on temperature, but this dependence is generally weak compared with the exponential factor, which is extremely sensitive to temperature. This exponential dependence on temperature is called *Arrhenius* behavior, and it is characteristic of physical processes that depend on thermal fluctuations to get over a barrier. Figure 8.15a shows a plot of the predicted viscosity as a function of temperature.

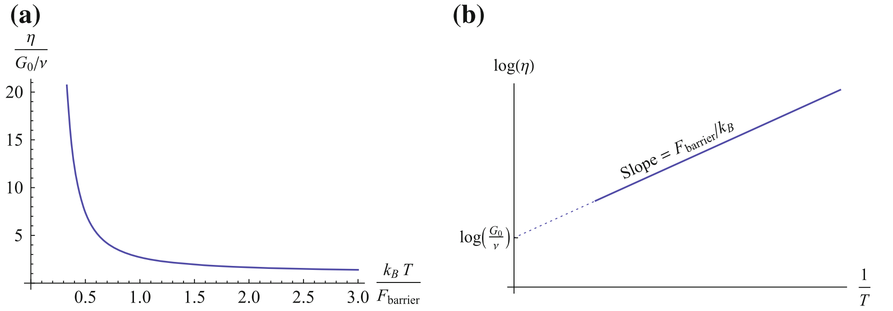


Fig. 8.15 Prediction for viscosity as a function of temperature. **a** Plotted on a linear scale. **b** Conventional Arrhenius plot of $\log(\eta)$ as a function of $1/T$

When the temperature is high, the viscosity is low. As the temperature decreases, the viscosity increases. This increase becomes especially sharp when $k_B T$ goes below F_{barrier}.

By the way, people who analyze data for viscosity as a function of temperature (or for any other thermally activated process) typically take the logarithm of both sides of Eq. (8.19) to obtain

$$\log(\eta) = \log\left(\frac{G_0}{\nu}\right) + \left(\frac{F_{\text{barrier}}}{k_B}\right)\left(\frac{1}{T}\right). \tag{8.20}$$

Inspired by this equation, they make an *Arrhenius plot* of $\log(\eta)$ as a function of $1/T$, as shown in Fig. 8.15b. Equation (8.20) predicts that this plot will be a straight line with slope of F_{barrier}/k_B and intercept of $\log(G_0/\nu)$, when extrapolated to $T^{-1} \to 0$ or $T \to \infty$. By fitting the experimental data, they can directly read off these two parameters.

8.6 Glass Transition

How well does the theory in the previous section compare with experiment? Suppose we cool a liquid from high to low temperature. We can consider two cases, depending on the cooling rate:

Case 1: Slow cooling

At high temperature, the viscosity follows the Arrhenius law of Eq. (8.19). Once the temperature reaches the phase transition from liquid to crystal, at a temperature T_{cryst}, the behavior changes. Nuclei of the crystal phase begin to form in the liquid. These nuclei grow, and eventually all of the liquid is transformed into the crystal phase. Once the system is in the crystal phase, it cannot flow viscously in response to a shear stress. Instead, it displays an elastic response to a shear stress, with a shear modulus.

This sequence of steps is the normal *equilibrium* behavior as the temperature decreases. In this sequence, there is plenty of time for nucleation and growth because the system is being cooled *slowly.*

Case 2: Fast cooling

At high temperature, the viscosity follows the Arrhenius law of Eq. (8.19), as in the previous case. Eventually, the temperature crosses below T_{cryst}, but now we are cooling so quickly that there is not enough time for nucleation and growth of the crystal phase. As a result, the system remains in the metastable liquid phase. This metastable phase is called a *supercooled liquid,* because it is cooled below the equilibrium crystallization point.

We can still measure the viscosity of the supercooled liquid as a function of temperature. For some range of temperature, it approximately follows the Arrhenius law of Eq. (8.19). However, when the temperature becomes low enough, the viscosity increases even more rapidly than predicted by this equation. Instead, it follows the *Vogel-Fulcher law:*

$$\eta = \eta_0 \exp\left(\frac{B}{T - T_0}\right), \tag{8.21}$$

where T_0 is some temperature below T_{cryst}, and η_0 and B are constants. (To be specific: At high temperature, the Arrhenius and Vogel-Fulcher predictions are basically identical. We can only tell the difference between Arrhenius and Vogel-Fulcher when the temperature becomes close to T_0, and then the viscosity follows Vogel-Fulcher.)

As the temperature continues to decrease, the viscosity increases tremendously. At some point, the viscosity becomes so high that the system can no longer flow on any experimental time scales. We might need to wait for years, or millions of years, to observe any flow. This change is called the *glass transition,* and it occurs at the temperature T_G. Note that $T_0 < T_G < T_{\text{cryst}}$: The glass transition is below the equilibrium crystallization transition, but it is above the point T_0 where the extrapolated viscosity goes to infinity.

The exact definition of the glass transition temperature T_G is a matter of convention. One common convention is that T_G is the temperature at which the viscosity grows to 10^{12} Pa s. This convention may seem arbitrary, but we should notice that the viscosity increases extremely rapidly as T approaches T_0. For this reason, if our criterion were a viscosity of 10^{13} or 10^{14} Pa s, we would report only a slightly different glass transition temperature.

Students sometimes ask where a glass should go on a phase diagram like Fig. 8.5 or 8.6. The answer is: Nowhere! A glass is not an equilibrium phase at any temperature or pressure. At low temperature, the equilibrium phase is a crystal. As far as we know, a glass is never the minimum of the free energy. Rather, a glass is a kinetic phenomenon, which occurs when a disordered material is stuck out of equilibrium. Even so, glasses are common solid materials, which occur both in nature and in technology. No discussion of solids can be complete without mentioning them.

You may notice that the Vogel-Fulcher law of Eq. (8.21) looks very similar to the scaling relations for physical properties near second-order phase transitions. Based on this similarity, some researchers have speculated that there might be a hidden phase transition to an *ideal glass*—not to be confused with an ideal gas! The transition to the ideal glass would occur at the temperature T_0, if only the system could remain in equilibrium at T_0, but actually it falls out of equilibrium.

Currently, the study of the glass transition is considered as part of a broader subject called *jamming,* which is the general study of systems that form rigid non-equilibrium structures. This field includes the study of granular materials like sand, which become jammed under shear stress, and even the study of traffic jams among cars. Since the late 1990s, this area has become an important part of materials research.

Further Reading

The structure and properties of crystals are discussed in books on solid-state physics. The most widely used introductory textbook is:

1. C. Kittel, *Introduction to Solid State Physics*, 8th edn. (Wiley, 2005)

A clear and concise discussion of elasticity, viscosity, and the glass transition is in:

2. R.A.L. Jones, *Soft Condensed Matter* (Oxford, 2002)

A further, more extensive discussion of the physics of glasses is in:

3. R. Zallen, *The Physics of Amorphous Solids* (Wiley, 1983)

Chapter 9
Second Mathematical Interlude: Tensors

Abstract Working with tensors is essential for describing any physical phenomena that depend on direction. This chapter begins by explaining the motivation to use tensors, so that physics will be independent of any arbitrary choice of coordinate system. It then shows how to work with tensors and relates tensor calculations to more familiar vector and matrix calculations. It concludes with suggestions about how to visualize tensors.

In the previous chapter, we saw that crystals have two types of order, positional and orientational. We discussed how to describe positional order, but we have not yet discussed how to describe orientational order. In order to describe orientational order, or any other properties of materials that depend on direction, we need a branch of mathematics called *tensors*.

In my experience, most science and engineering students are quite familiar with *vectors* and *matrices,*[1] but are not so familiar with tensors. In this chapter, I will present an introduction to tensors, so that we can use them later.

9.1 What is a Tensor?

To introduce the concept of tensors, let us consider how to write equations that are physically meaningful.

An important principle of physics must be: *Nature doesn't care about me.* The laws of nature cannot depend on any arbitrary choices that I might make.

One arbitrary choice is *units.* The laws of nature cannot depend on whether we are working in SI units, British Imperial units, or any other system. A student might say "The volume of my lunchbox is equal to the area of my laptop screen." This statement is totally meaningless! It might be numerically true if all lengths are measured in

[1] If you have not already studied vectors and matrices, you should learn about them before you read this chapter. One useful introduction to vectors, matrices, and other concepts of linear algebra is in the videos of the Khan Academy (http://www.khanacademy.org).

J.V. Selinger, *Introduction to the Theory of Soft Matter*,
Soft and Biological Matter, DOI 10.1007/978-3-319-21054-4_9

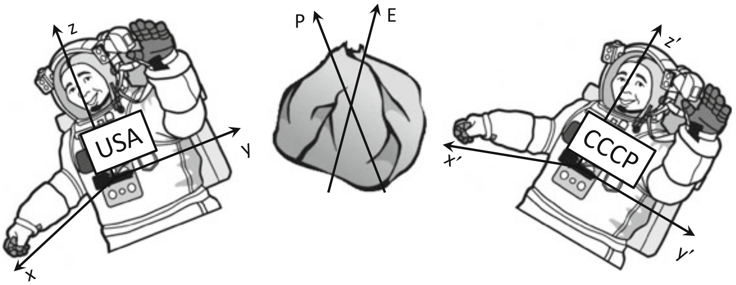


Fig. 9.1 Example of an American astronaut and a Russian cosmonaut observing the same rock on the International Space Station, each with his own Cartesian coordinate system

feet, but not if lengths are measured in any other units. This is the reason why science and engineering students need to learn to work with units correctly.

Another arbitrary choice is *coordinate system.* We can set up many different Cartesian coordinate systems, which are all rotated versions of each other. As an example, Fig. 9.1 shows an American astronaut and a Russian cosmonaut observing the same rock on the International Space Station. Each of them has his own Cartesian coordinate system. For this reason, they will report different numerical values for the x, y, and z components of the rock's position or velocity.

What can they agree on? In other words, what physically meaningful statements can they make, which do not depend on their individual coordinate systems?

First, they can agree on the *magnitudes* of vectors. Suppose they each calculate the magnitude of the velocity vector (squared),

$$|\boldsymbol{v}|^2 = \boldsymbol{v} \cdot \boldsymbol{v} = v_x^2 + v_y^2 + v_z^2. \tag{9.1}$$

You might worry that they could get different answers, because this expression depends on the vector components v_x, v_y, and v_z, which are different for the American and the Russian. However, thanks to the Pythagorean theorem, this sum of the squares of vector components turns out to be the same for both observers. Hence, the magnitude of the velocity vector is a *scalar,* a number which is *invariant* under rotation of the coordinate system. The same is true for the magnitude of any other vector. (By the way, note that they would *not* agree about the sum of fourth powers, $v_x^4 + v_y^4 + v_z^4$. The Pythagorean theorem only works for the sum of squares, not for any other power!)

Second, they can agree on *dot products.* Suppose they each measure the force vector $\boldsymbol{F}$ and the velocity vector $\boldsymbol{v}$, and calculate the power

$$\text{Power} = \boldsymbol{F} \cdot \boldsymbol{v} = F_x v_x + F_y v_y + F_z v_z. \tag{9.2}$$

Again, you might worry that they could get different answers, because this expression depends on the vector components of the force and velocity, which are different in the two coordinate systems. However, the dot product also turns out to be invariant under rotation of coordinates. You already knew this, because you recall that the dot product can be interpreted as $\boldsymbol{F} \cdot \boldsymbol{v} = |\boldsymbol{F}||\boldsymbol{v}| \cos\theta$, where θ is the angle between the vectors. Because the two observers have the same values of $|\boldsymbol{F}|$ and $|\boldsymbol{v}|$, and the same angle θ, they get the same dot product. Hence, the coordinate dependence of the vector components cancels out, and the dot product is a scalar. (In the previous case, $|\boldsymbol{v}|^2$ is just the dot product of $\boldsymbol{v}$ with itself. Hence, this is a just a special case of a dot product.)

Third, they can agree on *vector equations*. Suppose they each measure the force vector $\boldsymbol{F}$ and the acceleration vector $\boldsymbol{a}$, as well as the rock's mass m. They would each find that

$$\boldsymbol{F} = m\boldsymbol{a}, \tag{9.3}$$

or equivalently

$$\begin{aligned} F_x &= m a_x, \\ F_y &= m a_y, \\ F_z &= m a_z. \end{aligned} \tag{9.4}$$

This result is remarkable, because they would each measure different components for the force and acceleration vectors. It works because the force and acceleration vectors are not just any arbitrary lists of three numbers. Rather, the force vector $\boldsymbol{F}$ is a single 3D object, and the acceleration vector $\boldsymbol{a}$ is another 3D object. The vector Eq. (9.3) expresses a relationship between these 3D objects. Hence, we would say that this vector equation is an *invariant* equation, which does not change under rotation of coordinates.

Now let us consider a more interesting example. Suppose they want to investigate the dielectric properties of the rock. They apply an electric field $\boldsymbol{E}$, and measure the resulting electrostatic polarization $\boldsymbol{P}$. Because the rock has a complicated anisotropic crystal structure, the induced polarization is not necessarily parallel to the applied field. Instead, these vectors are related by

$$\boldsymbol{P} = \boldsymbol{\alpha} \cdot \boldsymbol{E} \tag{9.5}$$

or equivalently

$$\begin{pmatrix} P_x \\ P_y \\ P_z \end{pmatrix} = \begin{pmatrix} \alpha_{xx} & \alpha_{xy} & \alpha_{xz} \\ \alpha_{yx} & \alpha_{yy} & \alpha_{yz} \\ \alpha_{zx} & \alpha_{zy} & \alpha_{zz} \end{pmatrix} \begin{pmatrix} E_x \\ E_y \\ E_z \end{pmatrix} = \begin{pmatrix} \alpha_{xx} E_x + \alpha_{xy} E_y + \alpha_{xz} E_z \\ \alpha_{yx} E_x + \alpha_{yy} E_y + \alpha_{yz} E_z \\ \alpha_{zx} E_x + \alpha_{zy} E_y + \alpha_{zz} E_z \end{pmatrix}, \tag{9.6}$$

where the α_{ij} coefficients represent the polarizability of the rock. This is also an invariant equation! The American and the Russian measure different values for all of the $\boldsymbol{E}$ components, all of the $\boldsymbol{P}$ components, and all of the $\boldsymbol{\alpha}$ components, but the same equation still works for both of them. It works because $\boldsymbol{E}$, $\boldsymbol{P}$, and $\boldsymbol{\alpha}$ are not just arbitrary lists of numbers. Rather, each of them is a single 3D object.

You already know that $\boldsymbol{E}$ and $\boldsymbol{P}$ are vectors. The crucial point is: $\boldsymbol{\alpha}$ is another type of 3D object, called a *tensor*. A tensor is a mathematical object that expresses the relationship between vectors. In this case, $\boldsymbol{\alpha}$ is the polarizability tensor, and it expresses the relationship between $\boldsymbol{E}$ and $\boldsymbol{P}$, showing how the rock responds to electric fields with all possible orientations.

A vector is a 3D object with a magnitude and a direction. It can be represented by its components in a Cartesian coordinate system, but it is not just a list of numbers. Similarly, a tensor is a more complex 3D object with magnitudes and directions. It also can be represented by its components in a Cartesian coordinate system. However, we always want to remember that it is a 3D object, not just a list of numbers.

You might ask: How can we tell the difference between the components of a tensor and any other list of numbers? The answer is: Look at what happens when we rotate the coordinate system. The components of a tensor must transform in a specific way, which I will explain below. This is the reason why books about tensors emphasize transformation rules so much!

Before going on, I should briefly mention two points to avoid misunderstanding. First, I consistently describe tensors in 3D, because this is the most common dimensionality in physics. However, there is nothing special about 3D. The mathematical principles in this chapter can be applied to tensors in 2D, or 4D, or any other dimension. Second, I consistently describe tensor components in Cartesian coordinates. It is possible to describe tensor components in curvilinear coordinate systems (such as spherical or cylindrical coordinates), but that is a more advanced topic, beyond the scope of this book.

9.2 Working with Tensors

To understand all the possible ways of expressing tensors, let us begin with 3D vectors, such as the electric field $\boldsymbol{E}$ or polarization $\boldsymbol{P}$. Vectors are a special case of tensors, called *tensors of rank 1.* For a vector, we have several possible types of notation:

1. We can write an abstract vector as $\boldsymbol{v}$. People sometimes decorate the symbol as $\vec{\boldsymbol{v}}$ or $\underline{\boldsymbol{v}}$, but I will not do that in this book.
2. We can expand the vector in terms of basis vectors in a Cartesian coordinate system
$$\boldsymbol{v} = v_x \hat{\boldsymbol{x}} + v_y \hat{\boldsymbol{y}} + v_z \hat{\boldsymbol{z}}. \tag{9.7}$$
Here, $\hat{\boldsymbol{x}}$, $\hat{\boldsymbol{y}}$, and $\hat{\boldsymbol{z}}$ are unit vectors in the x, y, and z directions. The "hat" over the vectors indicates that they each have unit magnitude.
3. We can make a list of the components in a column or row,

$$\boldsymbol{v} = \begin{pmatrix} v_x \\ v_y \\ v_z \end{pmatrix} \quad \text{or} \quad \boldsymbol{v} = (v_x, v_y, v_z). \tag{9.8}$$

4. We can write a equation for the ith component of the vector. For example, the abstract vector equation $\boldsymbol{v} = 2\boldsymbol{u}$ can be written as

   $$v_i = 2u_i. \tag{9.9}$$

   This notation means that the equation is true for $i = x$, y, and z, so the equation is really three simultaneous equations. We might say that i is a "free" index that can represent any direction. In casual language, people look at this equation and say that "v_i is a vector," but it would be more precise to say that v_i is a vector component.

   We can see an important rule for working with vector components: If an index appears *once* in some term, it must appear once in every term on both sides of the equation. This index is free, so the equation must be true for all values of that index: x, y, or z. This rule is necessary so that the component equation can be a component-by-component version of a vector equation, which works in all Cartesian coordinate systems.

Next, let us consider a *tensor of rank 2,* which expresses the relationship between two vectors. An example is the polarizability tensor $\boldsymbol{\alpha}$, which expresses the relationship between $\boldsymbol{E}$ and $\boldsymbol{P}$. It can be expressed using the same type of notation:

1. We can write an abstract tensor as $\boldsymbol{T}$. People sometimes decorate the symbol as $\overleftrightarrow{\boldsymbol{T}}$ or $\underline{\underline{\boldsymbol{T}}}$, but I will not do that in this book.
2. We can expand the vector in terms of basis tensors in a Cartesian coordinate system:

   $$\begin{aligned} \boldsymbol{T} &= T_{xx}\hat{\boldsymbol{x}}\hat{\boldsymbol{x}} + T_{xy}\hat{\boldsymbol{x}}\hat{\boldsymbol{y}} + T_{xz}\hat{\boldsymbol{x}}\hat{\boldsymbol{z}} + T_{yx}\hat{\boldsymbol{y}}\hat{\boldsymbol{x}} + T_{yy}\hat{\boldsymbol{y}}\hat{\boldsymbol{y}} + T_{yz}\hat{\boldsymbol{y}}\hat{\boldsymbol{z}} \\ &\quad + T_{zx}\hat{\boldsymbol{z}}\hat{\boldsymbol{x}} + T_{zy}\hat{\boldsymbol{z}}\hat{\boldsymbol{y}} + T_{zz}\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}. \end{aligned} \tag{9.10}$$

   (People do not really use that notation in practice, because it is just too awkward, but they could use it.)
3. We can make a list of the components in a matrix:

   $$\boldsymbol{T} = \begin{pmatrix} T_{xx} & T_{xy} & T_{xz} \\ T_{yx} & T_{yy} & T_{yz} \\ T_{zx} & T_{zy} & T_{zz} \end{pmatrix}. \tag{9.11}$$

4. We can write a equation for the ij component of the tensor. For example, the abstract tensor equation $\boldsymbol{T} = 2\boldsymbol{S}$ can be written as

   $$T_{ij} = 2S_{ij}. \tag{9.12}$$

This notation means that the equation is true for $i = x, y$, and z and $j = x, y$, and z; so the equation is really nine simultaneous equations. We might say that i and j are free indices that can represent any directions. In casual language, people look at this equation and say that "T_{ij} is a tensor," but it would be more precise to say that T_{ij} is a tensor component.

Again, we have the same rule for free indices: If an index appears *once* in some term, it must appear once in every term on both sides of the equation, and the equation must be true for all values of that index: x, y, or z. This rule is necessary so that the component equation can be a component-by-component version of a tensor equation, which works in all Cartesian coordinate systems.

We can also define tensors of rank higher than two, which express the relationship between multiple vectors. Such tensors can be written in the abstract tensor notation as $\boldsymbol{C}$, or they can be written out in components such as C_{ijkl}. The rank of a tensor is equal to the number of indices, so the example of C_{ijkl} is a tensor of rank 4. When we are working in 3D, each index of the tensor can represent x, y, or z, and hence the tensor C_{ijkl} has $3^4 = 81$ components. In general, for a tensor of rank r in d dimensions, there are d^r components.

In the previous section, we saw two examples where we combine vectors or tensors by summing over an index. One example is a *dot product* of two vectors, also known as an *inner product*:

$$\boldsymbol{u} \cdot \boldsymbol{v} = u_x v_x + u_y v_y + u_z v_z = \sum_{i=1}^{3} u_i v_i. \tag{9.13}$$

The other example is *matrix multiplication* of a matrix (or tensor of rank 2) times a vector:

$$\boldsymbol{M} \cdot \boldsymbol{v} = \begin{pmatrix} M_{xx} & M_{xy} & M_{xz} \\ M_{yx} & M_{yy} & M_{yz} \\ M_{zx} & M_{zy} & M_{zz} \end{pmatrix} \begin{pmatrix} v_x \\ v_y \\ v_z \end{pmatrix} = \begin{pmatrix} M_{xx} v_x + M_{xy} v_y + M_{xz} v_z \\ M_{yx} v_x + M_{yy} v_y + M_{yz} v_z \\ M_{zx} v_x + M_{zy} v_y + M_{zz} v_z \end{pmatrix}, \tag{9.14}$$

or equivalently

$$(\boldsymbol{M} \cdot \boldsymbol{v})_i = \sum_{j=1}^{3} M_{ij} v_j. \tag{9.15}$$

These two examples involve summing over a tensor index that appears *exactly twice* in a term. This procedure is called *contraction*, or *contracting over an index.* It is a very important part of working with tensors, because it cancels the direction dependence associated with this index.

In tensor equations, contraction occurs so often that there is a compact shorthand for it, called the *Einstein summation convention:* Whenever a tensor index occurs exactly twice in a term, we automatically sum over that index $= x$, y, and z. We do not even bother to write down the summation sign; we just remember to sum over it.

With the Einstein summation convention, the dot product can be written as

$$\boldsymbol{u} \cdot \boldsymbol{v} = u_i v_i, \tag{9.16}$$

where the repeated index i is automatically summed over x, y, and z. Intuitively, we might say that contracting over the index i cancels the coordinate dependence in u_i and v_i. Hence, the dot product makes a scalar out of two vectors.

Similarly, with this convention, the product of a matrix and a vector can be written as

$$(\boldsymbol{M} \cdot \boldsymbol{v})_i = M_{ij} v_j. \tag{9.17}$$

This contraction cancels the coordinate dependence associated with j, and hence makes a single vector out of a matrix and a vector. Likewise, the product of two matrices is

$$(\boldsymbol{M} \cdot \boldsymbol{N})_{ik} = M_{ij} N_{jk}. \tag{9.18}$$

Here, contracting over the index j cancels the coordinate dependence associated with this index and makes a single matrix (tensor of rank 2).

To summarize the results of this section, we have three rules for working with tensor components:

1. If an index appears *once* in some term, the index is free. This index must appear once in every term on both sides of the equation.
2. If an index appears *twice* in some term, we automatically contract (sum over) the index. This index does not need to appear in other terms.
3. If an index appears *more than twice* in some term, we made a mistake. It must only appear once or twice!

These rules make it easy to write expressions that are invariant (do not depend on choice of coordinate system) and difficult to write expressions that are not invariant (depend on choice of coordinate system).

9.3 Standard Vector and Matrix Expressions

Besides dot products and matrix multiplication, you probably know several other vector and matrix expressions. It is possible to write all of them in tensor component notation. Here are several examples:

Identity matrix

In 3D, the identity matrix is

$$\boldsymbol{I} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{9.19}$$

The ij component of this matrix is commonly written as

$$I_{ij} = \delta_{ij}, \tag{9.20}$$

where δ_{ij} is the *Kronecker delta* defined as

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases} \tag{9.21}$$

The matrix equation

$$\boldsymbol{I} \cdot \boldsymbol{v} = \boldsymbol{v} \tag{9.22}$$

is then equivalent to

$$\delta_{ij} v_j = v_i. \tag{9.23}$$

Cross product of vectors

The cross product of two vectors is defined as

$$\boldsymbol{u} \times \boldsymbol{v} = \begin{pmatrix} u_y v_z - u_z v_y \\ u_z v_x - u_x v_z \\ u_x v_y - u_y v_x \end{pmatrix}. \tag{9.24}$$

It combines the two vectors to make a third vector. This combination is antisymmetric, meaning that $\boldsymbol{u} \times \boldsymbol{v} = -\boldsymbol{v} \times \boldsymbol{u}$.

To write the cross product in tensor component notation, we must first define the *Levi-Civita symbol*

$$\epsilon_{ijk} = \begin{cases} 1 & \text{if } ijk = xyz,\ yzx,\ \text{or } zxy \text{ (an even permutation of the indices)}, \\ -1 & \text{if } ijk = xzy,\ yxz,\ \text{or } zyx \text{ (an odd permutation of the indices)}, \\ 0 & \text{otherwise.} \end{cases} \tag{9.25}$$

The Levi-Civita symbol is completely antisymmetric, meaning that it changes sign if we exchange any two indices (for example $\epsilon_{ijk} = -\epsilon_{jik}$), but it keeps the same sign if we make an even number of exchanges (for example $\epsilon_{ijk} = \epsilon_{kij}$). In terms of the Levi-Civita symbol, the cross product is given by

$$(\boldsymbol{u} \times \boldsymbol{v})_i = \epsilon_{ijk} u_j v_k. \tag{9.26}$$

On the right side of this equation, the repeated indices j and k are automatically summed over x, y, and z. The index i is free, and it appears on both sides of the equation.

One useful identity involving the Levi-Civita symbol is

$$\epsilon_{ijk}\epsilon_{imn} = \delta_{jm}\delta_{kn} - \delta_{jn}\delta_{km}. \tag{9.27}$$

This identity means that the contraction $\epsilon_{ijk}\epsilon_{imn}$ is $+1$ if j matches k and m matches n, -1 if j matches n and k matches m, and 0 otherwise. Using this identity, we can, for example, simplify $\boldsymbol{a} \times (\boldsymbol{b} \times \boldsymbol{c})$:

$$\begin{aligned}
[\boldsymbol{a} \times (\boldsymbol{b} \times \boldsymbol{c})]_i &= \epsilon_{ijk} a_j (\boldsymbol{b} \times \boldsymbol{c})_k = \epsilon_{ijk} a_j \epsilon_{kmn} b_m c_n \\
&= \epsilon_{ijk}\epsilon_{kmn} a_j b_m c_n = \epsilon_{kij}\epsilon_{kmn} a_j b_m c_n \\
&= (\delta_{im}\delta_{jn} - \delta_{in}\delta_{jm}) a_j b_m c_n = \delta_{im}\delta_{jn} a_j b_m c_n - \delta_{in}\delta_{jm} a_j b_m c_n \\
&= a_j b_i c_j - a_j b_j c_i = b_i a_j c_j - c_i a_j b_j \\
&= b_i (\boldsymbol{a} \cdot \boldsymbol{c}) - c_i (\boldsymbol{a} \cdot \boldsymbol{b}) = [\boldsymbol{b}(\boldsymbol{a} \cdot \boldsymbol{c}) - \boldsymbol{c}(\boldsymbol{a} \cdot \boldsymbol{b})]_i\,.
\end{aligned} \tag{9.28}$$

This equation is the component-by-component version of the familiar vector identity

$$\boldsymbol{a} \times (\boldsymbol{b} \times \boldsymbol{c}) = \boldsymbol{b}(\boldsymbol{a} \cdot \boldsymbol{c}) - \boldsymbol{c}(\boldsymbol{a} \cdot \boldsymbol{b}). \tag{9.29}$$

Tensor product (outer product, dyad) of vectors

Apart from the dot product (which makes a scalar) and the cross product (which makes a vector), two vectors can also be combined to make a tensor of rank 2. This combination is called a tensor product or outer product and is written as $\boldsymbol{u} \otimes \boldsymbol{v}$. The same combination is also sometimes called a dyad and is written as $\boldsymbol{uv}$, with no symbol between the vectors. It is defined as the matrix of all possible products of Cartesian components:

$$\boldsymbol{u} \otimes \boldsymbol{v} = \boldsymbol{uv} = \begin{pmatrix} u_x v_x & u_x v_y & u_x v_z \\ u_y v_x & u_y v_y & u_y v_z \\ u_z v_x & u_z v_y & u_z v_z \end{pmatrix}. \tag{9.30}$$

In tensor component notation, this combination can be written very easily as

$$(\boldsymbol{u} \otimes \boldsymbol{v})_{ij} = (\boldsymbol{uv})_{ij} = u_i v_j. \tag{9.31}$$

In this expression, both indices i and j are free, and nothing is contracted.

Gradient, divergence, and curl

All of our notation for vectors applies equally well to derivatives of vector fields. The gradient operator is

$$\nabla = \begin{pmatrix} \partial/\partial x \\ \partial/\partial y \\ \partial/\partial z \end{pmatrix} = \begin{pmatrix} \partial_x \\ \partial_y \\ \partial_z \end{pmatrix}. \tag{9.32}$$

The symbols on the right are just a compact notation for writing derivatives with respect to x, y, and z. Hence, the gradient of a scalar field $f(\boldsymbol{r})$ is

$$\nabla f = \begin{pmatrix} \partial f/\partial x \\ \partial f/\partial y \\ \partial f/\partial z \end{pmatrix} = \begin{pmatrix} \partial_x f \\ \partial_y f \\ \partial_z f \end{pmatrix}. \tag{9.33}$$

In tensor component notation, it is written as

$$(\nabla f)_i = \partial_i f. \tag{9.34}$$

Likewise, the divergence of a vector field $\boldsymbol{v}(\boldsymbol{r})$ is

$$\nabla \cdot \boldsymbol{v} = \partial_i v_i, \tag{9.35}$$

and the curl is

$$(\nabla \times \boldsymbol{v})_i = \epsilon_{ijk} \partial_j v_k. \tag{9.36}$$

I should briefly mention one other notation that you might occasionally see in the literature: People sometimes put a subscript after a comma to indicate a derivative. Some examples are

$$\begin{aligned} v_{i,j} &= \partial_j v_i, \\ f_{,ij} &= \partial_i \partial_j f. \end{aligned} \tag{9.37}$$

Trace of a matrix

The trace of a matrix, or tensor of rank 2, is the sum of the elements along the diagonal:

$$\operatorname{Tr} \boldsymbol{M} = M_{xx} + M_{yy} + M_{zz}. \tag{9.38}$$

In tensor component notation, it can be written as

$$\mathrm{Tr}\,\boldsymbol{M} = M_{ii}. \tag{9.39}$$

In this expression, contracting over the index i cancels the coordinate dependence associated with that index—just as in any other contraction, such as a dot product or matrix multiplication. Hence, the trace makes a scalar, an invariant number, out of a tensor of rank 2.

In a similar way, we can take the trace of a product of two matrices

$$\mathrm{Tr}(\boldsymbol{M} \cdot \boldsymbol{N}) = M_{ij} N_{ji}, \tag{9.40}$$

or the trace of a power of a matrix

$$\mathrm{Tr}\,\boldsymbol{M}^2 = M_{ij} M_{ji}, \tag{9.41}$$

$$\mathrm{Tr}\,\boldsymbol{M}^3 = M_{ij} M_{jk} M_{ki}. \tag{9.42}$$

Notice that the trace of the identity matrix is

$$\mathrm{Tr}\,\boldsymbol{I} = \delta_{ii} = \delta_{xx} + \delta_{yy} + \delta_{zz} = 3, \tag{9.43}$$

because we are in 3D space.

Determinant of a matrix

The determinant of a matrix is a surprisingly awkward concept to express in tensor component notation. It can be done by constructing the appropriate combination of traces of powers of the matrix. In 3D, this construction is

$$\begin{aligned} \det \boldsymbol{M} &= \frac{1}{6}(\mathrm{Tr}\,\boldsymbol{M})^3 - \frac{1}{2}(\mathrm{Tr}\,\boldsymbol{M}^2)(\mathrm{Tr}\,\boldsymbol{M}) + \frac{1}{3}\mathrm{Tr}\,\boldsymbol{M}^3 \\ &= \frac{1}{6} M_{ii} M_{jj} M_{kk} - \frac{1}{2} M_{ij} M_{ji} M_{kk} + \frac{1}{3} M_{ij} M_{jk} M_{ki}. \end{aligned} \tag{9.44}$$

Like the trace, the determinant is a scalar, which is independent of the choice of coordinate system.

9.4 Transformation Under Rotation

Because vectors and tensors are 3D objects, their components must transform in specific ways when we rotate the coordinate system. In this section, we will derive the rules for transformation under rotation. With these transformation rules, we can show that the contracting over an index really does cancel the coordinate dependence associated with that index.

Let us begin with a vector (tensor of rank 1). A vector v can be expressed in terms of its components in the coordinate system (x, y, z) as

$$\boldsymbol{v} = v_x \hat{\boldsymbol{x}} + v_y \hat{\boldsymbol{y}} + v_z \hat{\boldsymbol{z}}. \tag{9.45}$$

It can also be expressed in terms of its components in a rotated coordinate system (x', y', z') as

$$\boldsymbol{v} = v'_x \hat{\boldsymbol{x}}' + v'_y \hat{\boldsymbol{y}}' + v'_z \hat{\boldsymbol{z}}'. \tag{9.46}$$

Of course, these two expressions represent the *same* vector. Hence, we can set them equal to each other:

$$v'_x \hat{\boldsymbol{x}}' + v'_y \hat{\boldsymbol{y}}' + v'_z \hat{\boldsymbol{z}}' = v_x \hat{\boldsymbol{x}} + v_y \hat{\boldsymbol{y}} + v_z \hat{\boldsymbol{z}}. \tag{9.47}$$

Suppose we want to calculate the primed components in terms of the unprimed components. We can take the dot product of each side of the equation with $\hat{\boldsymbol{x}}'$. In this dot product, we note that $\hat{\boldsymbol{x}}' \cdot \hat{\boldsymbol{x}}' = 1$ and $\hat{\boldsymbol{x}}' \cdot \hat{\boldsymbol{y}}' = \hat{\boldsymbol{x}}' \cdot \hat{\boldsymbol{y}}' = 0$, because the primed coordinate system is still a Cartesian coordinate system. Hence, we obtain

$$v'_x = (\hat{\boldsymbol{x}}' \cdot \hat{\boldsymbol{x}}) v_x + (\hat{\boldsymbol{x}}' \cdot \hat{\boldsymbol{y}}) v_y + (\hat{\boldsymbol{x}}' \cdot \hat{\boldsymbol{z}}) v_z. \tag{9.48}$$

Next we take the dot products of Eq. (9.47) with $\hat{\boldsymbol{y}}'$ and $\hat{\boldsymbol{z}}'$ to calculate v'_y and v'_z. We can put all three equations together into a single matrix equation:

$$\begin{pmatrix} v'_x \\ v'_y \\ v'_z \end{pmatrix} = \begin{pmatrix} \hat{\boldsymbol{x}}' \cdot \hat{\boldsymbol{x}} & \hat{\boldsymbol{x}}' \cdot \hat{\boldsymbol{y}} & \hat{\boldsymbol{x}}' \cdot \hat{\boldsymbol{z}} \\ \hat{\boldsymbol{y}}' \cdot \hat{\boldsymbol{x}} & \hat{\boldsymbol{y}}' \cdot \hat{\boldsymbol{y}} & \hat{\boldsymbol{y}}' \cdot \hat{\boldsymbol{z}} \\ \hat{\boldsymbol{z}}' \cdot \hat{\boldsymbol{x}} & \hat{\boldsymbol{z}}' \cdot \hat{\boldsymbol{y}} & \hat{\boldsymbol{z}}' \cdot \hat{\boldsymbol{z}} \end{pmatrix} \begin{pmatrix} v_x \\ v_y \\ v_z \end{pmatrix}. \tag{9.49}$$

Using the Einstein summation convention, this equation can be written compactly as

$$v'_i = R_{ij} v_j, \tag{9.50}$$

where $\boldsymbol{R}$ is the *rotation matrix*

$$\boldsymbol{R} = \begin{pmatrix} \hat{\boldsymbol{x}}' \cdot \hat{\boldsymbol{x}} & \hat{\boldsymbol{x}}' \cdot \hat{\boldsymbol{y}} & \hat{\boldsymbol{x}}' \cdot \hat{\boldsymbol{z}} \\ \hat{\boldsymbol{y}}' \cdot \hat{\boldsymbol{x}} & \hat{\boldsymbol{y}}' \cdot \hat{\boldsymbol{y}} & \hat{\boldsymbol{y}}' \cdot \hat{\boldsymbol{z}} \\ \hat{\boldsymbol{z}}' \cdot \hat{\boldsymbol{x}} & \hat{\boldsymbol{z}}' \cdot \hat{\boldsymbol{y}} & \hat{\boldsymbol{z}}' \cdot \hat{\boldsymbol{z}} \end{pmatrix}. \tag{9.51}$$

One common example is a rotation by an angle θ about the z-axis. In this case, the rotation matrix is just

$$\boldsymbol{R} = \begin{pmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{9.52}$$

There are similar simplifications for a rotation about the x- or y-axis. However, for an arbitrary rotation axis, the general form of $\boldsymbol{R}$ is the matrix of dot products.

Working from Eq. (9.47), we can also take dot products with $\hat{\boldsymbol{x}}$, $\hat{\boldsymbol{y}}$, and $\hat{\boldsymbol{z}}$ to find the unprimed components in terms of the primed components. The result is

$$\begin{pmatrix} v_x \\ v_y \\ v_z \end{pmatrix} = \begin{pmatrix} \hat{\boldsymbol{x}} \cdot \hat{\boldsymbol{x}}' & \hat{\boldsymbol{x}} \cdot \hat{\boldsymbol{y}}' & \hat{\boldsymbol{x}} \cdot \hat{\boldsymbol{z}}' \\ \hat{\boldsymbol{y}} \cdot \hat{\boldsymbol{x}}' & \hat{\boldsymbol{y}} \cdot \hat{\boldsymbol{y}}' & \hat{\boldsymbol{y}} \cdot \hat{\boldsymbol{z}}' \\ \hat{\boldsymbol{z}} \cdot \hat{\boldsymbol{x}}' & \hat{\boldsymbol{z}} \cdot \hat{\boldsymbol{y}}' & \hat{\boldsymbol{z}} \cdot \hat{\boldsymbol{z}}' \end{pmatrix} \begin{pmatrix} v'_x \\ v'_y \\ v'_z \end{pmatrix}. \tag{9.53}$$

We can see that the rotation matrix for this inverse transformation is $\boldsymbol{R}^T$, the transpose of the original rotation matrix. In terms of tensor components, we have

$$v_i = R^T_{ij} v'_j = R_{ji} v'_j = v'_j R_{ji}, \tag{9.54}$$

using the definition of the transpose $R^T_{ij} = R_{ji}$. By combining Eqs. (9.50) and (9.54), we see that

$$v_i = R^T_{ij} R_{jk} v_k \tag{9.55}$$

for any vector $\boldsymbol{v}$. This is only possible if

$$R^T_{ij} R_{jk} = R_{ji} R_{jk} = \delta_{ik}, \tag{9.56}$$

or equivalently

$$\boldsymbol{R}^T \boldsymbol{R} = \boldsymbol{I}, \qquad \boldsymbol{R}^T = \boldsymbol{R}^{-1}. \tag{9.57}$$

In other words, the transpose of the rotation matrix is its inverse. The term for a matrix that satisfies this relationship is an *orthogonal* matrix. You can verify explicitly that the special form of Eq. (9.52) is orthogonal.

Now suppose we want to calculate the dot product of two vectors. In the unprimed coordinate system, the dot product is

$$\boldsymbol{u} \cdot \boldsymbol{v} = u_i v_i = u_x v_x + u_y v_y + u_z v_z. \tag{9.58}$$

In the primed coordinate system, it is

$$\begin{aligned} \boldsymbol{u} \cdot \boldsymbol{v} &= u'_i v'_i = u'_x v'_x + u'_y v'_y + u'_z v'_z \\ &= R_{ij} u_j R_{ik} v_k = R_{ij} R_{ik} u_j v_k = \delta_{jk} u_j v_k \\ &= u_j v_j = u_x v_x + u_y v_y + u_z v_z. \end{aligned} \tag{9.59}$$

This calculation shows explicitly that the dot product is the same in the primed and unprimed coordinate systems, i.e., the dot product is a scalar, invariant under rotations of coordinates. It confirms that contracting over an index cancels the coordinate dependence associated with this index.

For a tensor of rank 2, the transformation rule is slightly more complicated, because we must transform both indices. We can write the rule in matrix form as

$$\begin{pmatrix} T'_{xx} & T'_{xy} & T'_{xz} \\ T'_{yx} & T'_{yy} & T'_{yz} \\ T'_{zx} & T'_{zy} & T'_{zz} \end{pmatrix} = \boldsymbol{R} \cdot \begin{pmatrix} T_{xx} & T_{xy} & T_{xz} \\ T_{yx} & T_{yy} & T_{yz} \\ T_{zx} & T_{zy} & T_{zz} \end{pmatrix} \cdot \boldsymbol{R}^T, \tag{9.60}$$

or equivalently in component form as

$$T'_{ij} = R_{im} T_{mn} R^T_{nj} = R_{im} R_{jn} T_{mn}. \tag{9.61}$$

This rule shows that the matrix of unprimed components must be sandwiched between $\boldsymbol{R}$ and $\boldsymbol{R}^T$, to obtain the matrix of primed components. As a consistency check, suppose we have a tensor of rank 2 multiplying a vector, $\boldsymbol{u} = \boldsymbol{T} \cdot \boldsymbol{v}$, so that

$$u_i = T_{ij} v_j. \tag{9.62}$$

In the primed coordinate system, the product becomes

$$u'_i = T'_{ij} v'_j = R_{im} T_{mn} R^T_{nj} R_{jk} v_k = R_{im} T_{mn} \delta_{nk} v_k = R_{im} T_{mn} v_n = R_{im} u_m. \tag{9.63}$$

Hence, the components of $\boldsymbol{u}$ transform in the correct way for a vector, and our notation is consistent.

For a tensor of higher rank, we just have to apply a rotation matrix for each index. With the fourth-rank tensor C_{ijkl}, for example, we have

$$C'_{ijkl} = R_{ia} R_{jb} R_{kc} R_{ld} C_{abcd}. \tag{9.64}$$

With this rule, we can see that all tensor expressions transform appropriately, so that tensor equations are invariant under rotation of coordinates.

9.5 Transformation Under Inversion

So far we have only discussed how tensor components transform under *rotation* of coordinates. We should also consider how they transform under *inversion*. This transformation is especially important for cross products, as well as other tensor expressions that include the Levi-Civita symbol. It can be explained in two different ways, which will turn out to be equivalent:

9.5.1 Right-Hand Rule Versus Left-Hand Rule

At the beginning of this chapter, I argued that the laws of physics cannot depend on any arbitrary choices that I make. One arbitrary choice is units, and a second arbitrary choice is coordinate system. A third arbitrary choice is the right-hand rule. Whenever we construct a cross product, we use the right-hand rule. What is so special about the right hand? Why not use the left-hand rule instead?

Earlier in this chapter, we saw that the cross product can be defined in terms of the Levi-Civita symbol ϵ_{ijk}. The same arbitrary choice that goes into the cross product also goes into ϵ_{ijk}. We picked a definition where $\epsilon_{xyz} = \epsilon_{yzx} = \epsilon_{zxy} = 1$ and $\epsilon_{xzy} = \epsilon_{yxz} = \epsilon_{zyx} = -1$. We could have made the opposite choice.

We can categorize tensor expressions based on whether they depend on this arbitrary choice of right-hand rule versus left-hand rule:

- If an expression does *not* involve ϵ_{ijk}, then it does *not* depend on this arbitrary choice. Likewise, if an expression involves an *even* power of ϵ_{ijk}, then it does *not* depend on this arbitrary choice. In this case, it is called a *proper* scalar, a *proper* vector (also known as a *polar* vector), or a *proper* tensor.
- By contrast, if an expression involves ϵ_{ijk} to the first power, or to any *odd* power, then it depends on this arbitrary choice. In this case, it is called a *pseudoscalar, pseudovector* (also known as an *axial* vector), or a *pseudotensor.*

Here are some physical examples:

- The position vector $\boldsymbol{r}$ and linear momentum $\boldsymbol{p}$ are proper vectors. The angular momentum $\boldsymbol{L} = \boldsymbol{r} \times \boldsymbol{p}$ is a pseudovector, because it has one power of ϵ_{ijk} in the cross product.
- The force $\boldsymbol{F}$ is a proper vector. The torque $\boldsymbol{\tau} = \boldsymbol{r} \times \boldsymbol{F}$ is a pseudovector.
- The electric and magnetic fields induced by a moving point charge are given by

$$\begin{aligned} \boldsymbol{E} &= \frac{q}{4\pi\epsilon_0}\frac{\hat{\boldsymbol{r}}}{r^2}, \\ \boldsymbol{B} &= \frac{\mu_0 q\boldsymbol{v}}{4\pi} \times \frac{\hat{\boldsymbol{r}}}{r^2}. \end{aligned} \tag{9.65}$$

  The electric field is a proper vector, and the magnetic field is a pseudovector.
- If a point charge moves in electric and magnetic fields, it experiences a Lorentz force

$$\boldsymbol{F} = q(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}). \tag{9.66}$$

  Both terms in this Lorentz force are proper vectors. (The second term is a proper vector because it is the proper vector $\boldsymbol{v}$ crossed with the pseudovector $\boldsymbol{B}$, and hence it has two powers of ϵ_{ijk}.)

9.5.2 Inversion of Coordinates

As an alternative way to think about vectors and pseudovectors, we will stick with the right-hand rule and invert the coordinate system. Inverting the coordinate system means changing $x \rightarrow -x$, $y \rightarrow -y$, and $z \rightarrow -z$. Under this transformation, all proper scalars remain unchanged. For a proper vector, all components change sign, so that $v_i \rightarrow -v_i$. For a proper tensor of rank 2, all components remain unchanged. This is reasonable, because a proper second-rank tensor expresses the relationship between two proper vectors, and the sign changes for those vector components will cancel. In general, under a coordinate inversion, all components of even-rank tensors remain unchanged, and all components of odd-rank tensors change sign.

What about a cross product, such as $\boldsymbol{w} = \boldsymbol{u} \times \boldsymbol{v}$? The components of a cross product are defined by $w_x = u_y v_z - u_z v_y$, and so forth. If the u components and the v components all change sign, these sign changes will cancel, and hence the w components will remain unchanged. Hence, under the inversion, the behavior of $\boldsymbol{w}$ is the opposite of the behavior of a proper vector. This is an alternative way to see that $\boldsymbol{w}$ is a pseudovector.

The same argument can be made for the Levi-Civita symbol itself. The cross product components are defined in terms of the Levi-Civita components as $w_i = \epsilon_{ijk} u_j v_k$. If the w components remain unchanged under inversion, and the u and v components both change sign, then the ϵ_{ijk} components must remain the same under inversion. Although the components of a proper third-rank tensor change sign under inversion, the components of the Levi-Civita symbol have the opposite behavior; they remain unchanged. This is a way to see that the Levi-Civita symbol is pseudotensor.

All of the examples of proper vectors and pseudovectors from the previous subsection still apply here. Inverting the coordinate system is just another way to describe the same mathematical behavior as switching from the right-hand rule to the left-hand rule.

9.5.3 Reflection of Coordinates

You might ask: What if we do not change the sign for all three coordinates x, y, and z, but only some of them?

First suppose we change the sign for two of the coordinates. For example, suppose we change $x \rightarrow -x$ and $y \rightarrow -y$, but keep $z \rightarrow +z$. This change is just a rotation through an angle of 180° about the z-axis. Indeed, if we take the rotation matrix of Eq. (9.52) and substitute $\theta = 180°$, we obtain

$$\boldsymbol{R} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{9.67}$$

which represents changing the signs of x and y, but not z.

Now suppose we change the sign of just one coordinate. For example, suppose we keep $x \to +x$ and $y \to +y$, but change $z \to -z$. This change is a *reflection* in the xy plane. The important point is that this reflection is related to the inversion and the $\theta = 180°$ rotation. We can write

$$\text{inversion} = (\text{rotation of } 180° \text{ about } z) \cdot (\text{reflection in } xy \text{ plane}), \tag{9.68}$$

or equivalently,

$$(\text{reflection in } xy \text{ plane}) = (\text{rotation of } 180° \text{ about } z) \cdot \text{inversion}. \tag{9.69}$$

Hence, the distinction between proper expressions and pseudo-expressions applies to reflections as well as inversions.

9.6 Magnitudes, Directions, and Visualization

When we work with vectors, we have a choice: We can represent them by Cartesian components, or we can represent them by their magnitudes and directions. These two descriptions are equivalent to each other, and they have the same number of degrees of freedom. In 3D, a vector clearly has three Cartesian components. Similarly, it has one magnitude, and its direction is described by two angles (which might be θ and ϕ in spherical coordinates, or the latitude and longitude on a globe). In general, the description in terms of Cartesian components is more convenient for calculations, but the description in terms of magnitude and direction is more convenient for visualization—it is easy to image a vector as an arrow with some length, pointing in some direction.

You might ask: What about tensors? Do they have magnitudes and directions? If so, can we use these magnitudes and directions to visualize tensors? The answer is yes—tensors have magnitudes and directions, but we have to do a little work to extract them! In this section, I will show how to extract the magnitudes and directions for tensors of rank 2, and I will present a method for visualizing them. (Generalization to higher-rank tensors is a further challenge.)

Suppose we have a tensor $\boldsymbol{T}$, with components T_{ij} in a Cartesian coordinate system. This tensor has nine independent degrees of freedom. The first step is to break it into symmetric and antisymmetric parts. Any arbitrary second-rank tensor can be written as

$$T_{ij} = S_{ij} + A_{ij}, \tag{9.70}$$

where

$$S_{ij} = \frac{T_{ij} + T_{ji}}{2}, \qquad A_{ij} = \frac{T_{ij} - T_{ji}}{2}. \tag{9.71}$$

Here, $\boldsymbol{S}$ is a symmetric tensor, meaning that $S_{ij} = S_{ji}$, and it has six degrees of freedom. Likewise, $\boldsymbol{A}$ is an antisymmetric tensor, meaning that $A_{ij} = -A_{ji}$, and it has three degrees of freedom. We can now try to analyze the symmetric and antisymmetric parts separately.

Antisymmetric part

Because of the antisymmetry, $\boldsymbol{A}$ can be related to a pseudovector $\boldsymbol{P}$ by

$$P_k = \epsilon_{ijk} A_{ij}, \qquad A_{ij} = \frac{1}{2}\epsilon_{ijk} P_k. \tag{9.72}$$

Here, $\boldsymbol{P}$ is a pseudovector because it has one factor of ϵ_{ijk}, while $\boldsymbol{A}$ is a proper second-rank tensor. As a pseudovector, $\boldsymbol{P}$ has three degrees of freedom, the same as the antisymmetric tensor $\boldsymbol{A}$, so we have correctly accounted for the degrees of freedom. Physically, we would say that $\boldsymbol{A}$ represents a circulation about $\boldsymbol{P}$, with a handedness given by the right-hand rule. For example, $\boldsymbol{A}$ might represent a flow of material, while $\boldsymbol{P}$ represents the angular momentum. Alternatively, $\boldsymbol{A}$ might represent an electric current loop, while $\boldsymbol{P}$ represents the magnetic field.

Because $\boldsymbol{P}$ is a pseudovector, we know that it has one magnitude $|\boldsymbol{P}|$, as well as a direction given by two angles. We can say that the magnitude and direction of $\boldsymbol{A}$ are represented by the magnitude and direction of $\boldsymbol{P}$. Furthermore, we know how to visualize $\boldsymbol{P}$ as an arrow in 3D. We can say that $\boldsymbol{A}$ is represented by the same arrow in 3D, with the understanding that $\boldsymbol{A}$ is really a circulation around that arrow.

Symmetric part

Because S_{ij} is a real symmetric matrix, we can diagonalize it to find the eigenvalues (α, β, γ) and corresponding eigenvectors $(\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}, \hat{\boldsymbol{c}})$. The eigenvectors are three unit vectors, perpendicular to each other, which represent the principal axes of $\boldsymbol{S}$. To describe the orientations of these eigenvectors, we need three angles (two angles for $\hat{\boldsymbol{a}}$, one angle for $\hat{\boldsymbol{b}}$ because it is perpendicular to $\hat{\boldsymbol{a}}$, and no angles for $\hat{\boldsymbol{c}}$ because it is perpendicular to $\hat{\boldsymbol{a}}$ and $\hat{\boldsymbol{b}}$). The eigenvalues are scalars that represent the magnitude of the tensor associated with each eigenvector. Indeed, we can write the symmetric tensor as

$$\boldsymbol{S} = \alpha\hat{\boldsymbol{a}}\hat{\boldsymbol{a}} + \beta\hat{\boldsymbol{b}}\hat{\boldsymbol{b}} + \gamma\hat{\boldsymbol{c}}\hat{\boldsymbol{c}}, \tag{9.73}$$

or equivalently

$$S_{ij} = \alpha a_i a_j + \beta b_i b_j + \gamma c_i c_j. \tag{9.74}$$

Hence, we can say that $\boldsymbol{S}$ has three magnitudes, and an orientation given by three angles. This adds up to six, which is the correct number of degrees of freedom for $\boldsymbol{S}$.

Combining the symmetric and antisymmetric parts, the whole tensor $\boldsymbol{T}$ has four magnitudes, and an orientation given by five angles, for a total of nine degrees of freedom.

To visualize $\boldsymbol{S}$, an ideal solution would be: Draw an ellipsoid in 3D with the principal axes given by $\hat{\boldsymbol{a}}$, $\hat{\boldsymbol{b}}$, and $\hat{\boldsymbol{c}}$, and with radii along these axes given by α, β, γ. This almost works! There is just one problem: The eigenvalues of a real symmetric matrix are guaranteed to be real, but they are not guaranteed to be positive. There is a risk that some of the eigenvalues might be negative. In this case, we would face the problem of how to draw an ellipsoid with a negative radius.

One possible solution to this problem is to add a positive offset ω to the eigenvalues. The radii of the ellipsoid along its principal axes can then be $(\alpha+\omega)$, $(\beta+\omega)$, and $(\gamma+\omega)$. The question then is how to choose ω. If ω is too small, one of the radii might be negative. If ω is too large, the ellipsoid will just look like a sphere.

My own personal proposal is to choose the offset ω so that the product of radii is

$$(\alpha+\omega)(\beta+\omega)(\gamma+\omega) = 1. \tag{9.75}$$

The advantage of this choice is that the *volume* of the ellipsoid is fixed. At fixed volume, the ellipsoid can be more or less eccentric, with extension along different axes. The disadvantage of this choice is that it throws away information about the average eigenvalue, i.e., the *isotropic* part of the tensor. Indeed, if we add an extra isotropic part $S_{ij} \to S_{ij} + c\delta_{ij}$, then the offset changes by $\omega \to \omega - c$, and hence the ellipsoid remains exactly the same. Hence, with this offset, the visualization does not really represent all six degrees of freedom in $\boldsymbol{S}$, but only five degrees of freedom for the anisotropic part of $\boldsymbol{S}$. The isotropic part (the average eigenvalue) must be reported separately.

Figure 9.2 shows an example of a tensor visualized through this proposed method. In this visualization, there are three separate sections: The isotropic part of the tensor is represented by a sphere of unit radius, with an inscribed number specifying the average eigenvalue.[2] The anisotropic symmetric part of the tensor is represented by an ellipsoid, with the appropriate principal axes and radii (with offset). The antisymmetric part of the tensor is represented by the corresponding pseudovector, with the understanding that it really means circulation around that arrow.

The interactive version of this figure allows you to enter your own tensor and see the visualization. I strongly encourage you to try it!

Further Reading

Tensors are discussed in the same two textbooks listed at the end of Chap. 5. A more introductory treatment of tensors is in:

1. G.B. Arfken, H.J. Weber, F.E. Harris, *Mathematical Methods for Physicists: A Comprehensive Guide*, 7th edn. (Elsevier, 2013)

[2] One might want the radius of the sphere to be the average eigenvalue, but the average eigenvalue might be negative.

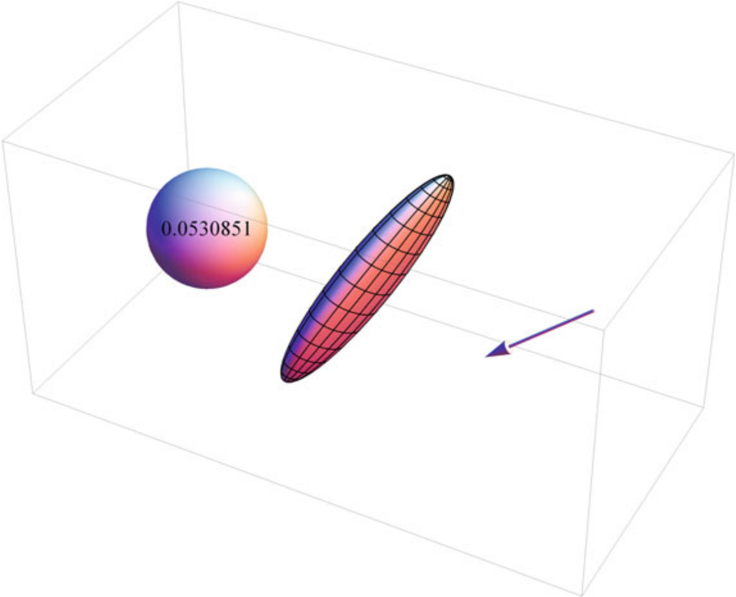


Fig. 9.2 Proposed method for visualizing a tensor (with a total of nine degrees of freedom). The visualization has three parts: isotropic (one degree of freedom), anisotropic symmetric (five degrees of freedom), and antisymmetric (three degrees of freedom) (Interactive version at http://www.springer.com/cda/content/document/cda_downloaddocument/Selinger+Interactive+Figures.zip?SGWID=0-0-45-1509169-p177545420.)

A more advanced treatment is in:

2. M. Stone, P. Goldbart, *Mathematics for Physics: A Guided Tour for Graduate Students* (Cambridge, 2009)

In particular, the latter book emphasizes the distinction between covariant and contravariant indices, which is essential in curvilinear coordinates but not in Cartesian coordinates. A pre-publication version of that book is available from the author's website http://www.physics.gatech.edu/~pgoldbart6/PG_MS_MfP.htm.

Chapter 10
Liquid Crystals

Abstract Liquid crystals are phases with more order than liquids but less order than crystals. This chapter presents a detailed discussion of the most common liquid-crystal phase, the nematic phase, and then more briefly discusses the cholesteric and other liquid-crystal phases. It begins by considering the orientational order of the nematic phase and constructing the nematic order parameter, which is a tensor. It goes on to discuss the physical mechanisms that control the magnitude and direction of nematic order. For the magnitude, it presents the Landau-de Gennes, Maier-Saupe, and Onsager theories of the isotropic-nematic transition. For the direction, it presents the Frank free energy, and uses this free energy for calculations of surface anchoring and nematic defects.

One fascinating part of the theory of soft matter is the study of *liquid crystals*—a subject that brings together all of the physical and mathematical concepts developed in previous chapters. This chapter provides an introduction to the theory of liquid crystals. It mainly discusses the nematic phase, which is the most common liquid-crystal phase, and then briefly considers cholesteric and other liquid-crystal phases at the end.

10.1 Order and Symmetry

In Chap. 8, we saw that there are two important differences between liquids and crystals from the perspective of order and symmetry:

(1) In liquids, all positions are equivalent; an atom is equally likely to be in any position. By contrast, crystals have certain *special positions*—the lattice sites—where there is an enhanced probability of finding an atom. Liquids are *uniform* but crystals are *nonuniform*. This distinction can be seen macroscopically as sharp peaks in the X-ray diffraction from crystals, but not from liquids. For that reason, we say that crystals have less symmetry than liquids; the perfect positional symmetry of a liquid is broken in a crystal. Equivalently, we say that crystals have *positional order*—the choice of the special positions is a type of order that occurs in crystals but not in liquids.

J.V. Selinger, *Introduction to the Theory of Soft Matter*,
Soft and Biological Matter, DOI 10.1007/978-3-319-21054-4_10

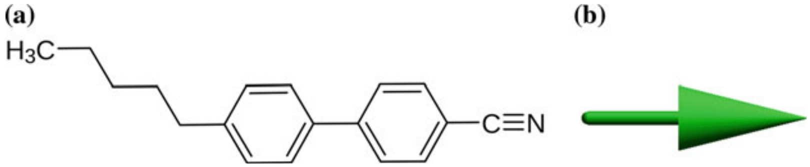


Fig. 10.1 **a** Molecular structure of 4-cyano-4′-pentylbiphenyl (5CB). From http://en.wikipedia.org/wiki/4-Cyano-4%27-pentylbiphenyl. **b** Coarse-grained representation of the molecular orientation as an *arrow*

(2) In liquids, all directions are equivalent; we are equally likely to find atoms along any direction. By contrast, crystals have certain *special directions*—the crystalline axes—where there is an enhanced probability of finding rows of atoms. Liquids are *isotropic* but crystals are *anisotropic*. This distinction can be seen macroscopically in, for example, the optical properties of crystals, which depend on the orientation. For that reason, we say that crystals have less symmetry than liquids; the perfect orientational symmetry of a liquid is broken in a crystal. Equivalently, we say that crystals have *orientational order*—the choice of the special directions is another type of order that occurs in crystals but not in liquids.

You might ask: Do these two types of order always go together? Is it ever possible to have one type of order, but not the other?

The answer to this question is: It is possible to have orientational order without positional order. There is a class of phases, called *liquid crystals* or *liquid-crystal phases* or *mesophases*, which are intermediate between liquids and crystals in the sense that they have partial order. Some of these phases have orientational order but no positional order. Other phases have orientational order as well as positional order in one or two directions, but not in all three dimensions. The field of liquid crystals has become an important part of modern science and technology, and it is the topic of this chapter.[1]

How is it possible to have a phase with orientational order but not positional order? The basic concept is that the phase is composed of molecules that are not spheres. Rather, the molecules are long and narrow, with some rigidity. In that case, the positions of the molecular centers of mass can be random, but the molecules can still tend to align with their neighbors. They can then form a phase with long-range order in the orientation of molecular alignment.

As an example, Fig. 10.1a shows a molecule of 4-cyano-4′-pentylbiphenyl, commonly called 5CB, which is one of the most widely used compounds that form a liquid-crystal phase. It has a central core with two benzene rings, which is fairly rigid, connected to a more flexible hydrocarbon chain. On a coarse-grained basis, we can neglect the detailed molecular structure and represent the entire molecule by

[1] By comparison, it is not possible to have positional order without at least some orientational order. Positional order necessarily requires some crystalline axes, which select certain directions in space. However, in the field of *molecular crystals,* there can be transitions among phases with different types of orientational order. This is a more specialized topic than liquid crystals, and I will not discuss it further in this book.

an arrow, as shown in Fig. 10.1b. We can then ask: When we have a sample of *many* such molecules, what is the statistical distribution of molecular (arrow) orientations?

Figure 10.2a shows a schematic illustration of the distribution of orientations in an ordinary liquid. We can see that the arrows are equally likely to point in any direction in 3D space. Hence, this ordinary liquid is called the *isotropic phase* of these molecules.

In a liquid-crystal phase, you might guess that the distribution would have the form shown in Fig. 10.2b. *However, this guess is generally not correct!* Fig. 10.2b shows a phase in which the molecules are aligned in some direction, which is $+\hat{z}$ in this example. Clearly, there are fluctuations in the orientation, but $+\hat{z}$ is the average. If such a phase existed, it would be called a *polar phase.* It would have the same type of order as a ferromagnet (similar to the Ising model discussed in Chap. 2, but with magnetic order that can align at any direction in 3D, not just up or down). However, this is not the liquid-crystal phase that normally forms in experiments!

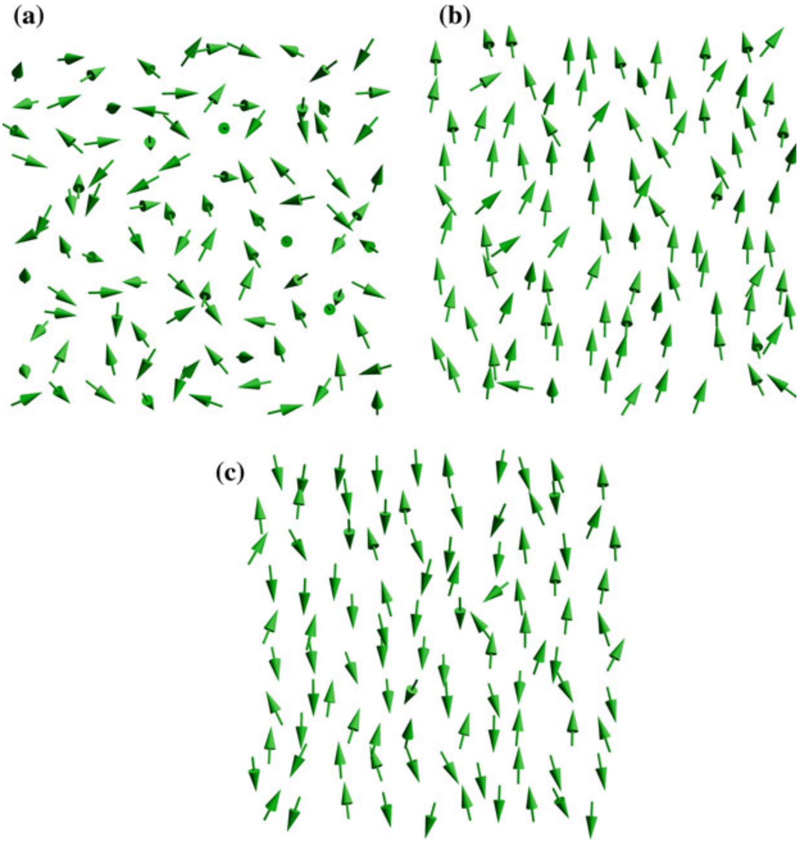


Fig. 10.2 Schematic illustrations of the distributions of molecular orientations. **a** Isotropic phase. **b** Polar phase. **c** Nematic phase

Instead, the most common liquid-crystal phase is the *nematic phase,* shown schematically in Fig. 10.2c. In the nematic phase, the molecules are aligned along some axis, which is $\pm\hat{z}$ in this example. They are equally likely to point up or down along the axis. Again, there are fluctuations in the orientation, but the $\pm\hat{z}$ axis is the average.

Among these three phases, the isotropic phase has the most symmetry (least order), because all directions are equivalent. The nematic phase has less symmetry (more order), because it has selected a special axis, $\pm\hat{z}$ in this example. The polar phase has the least symmetry (the most order), because it has selected a special direction along that axis, $+\hat{z}$ in this example.

The next task will be to define a mathematical approach to distinguish among the isotropic, nematic, and polar phases. We will do that by defining order parameters in the next section.

10.2 Nematic Order Parameter

When we discussed crystals in Chap. 8, we defined crystalline order parameters, which represent the magnitude and direction of the positional symmetry breaking. At that point, we did not need an orientational order parameter, because the crystalline wavevectors already contain information about the orientations of the crystalline axes. However, when we discuss liquid crystals, we need to describe orientational order *without* positional order. Hence, we now need to construct a orientational order parameter for the nematic liquid-crystal phase.

To construct an orientational order parameter, we must calculate an average over the orientations of all the molecules in an ensemble, as shown in Fig. 10.2a, b, or c. Hence, we represent the orientation of each molecule as a unit vector $\hat{\boldsymbol{\ell}}$. This unit vector has components ℓ_α, with $\alpha = x, y$, or z.

As a first try, we might construct an orientational order parameter by averaging all the unit vectors in the ensemble, which gives

$$\boldsymbol{M} = \langle \hat{\boldsymbol{\ell}} \rangle, \tag{10.1}$$

or equivalently in terms of components

$$M_\alpha = \langle \ell_\alpha \rangle. \tag{10.2}$$

This order parameter is quite analogous to the Ising order parameter of Chap. 2, except that it is a vector that can point in any direction in 3D space, not just up or down. In the isotropic phase of Fig. 10.2a, $\boldsymbol{M}$ averages to zero. In the polar phase of Fig. 10.2b, $\boldsymbol{M}$ has some nonzero average. What about the nematic phase of Fig. 10.2c? Because the molecules are equally likely to point up or down along the main axis, $\boldsymbol{M}$ averages to zero in the nematic phase, just as in the isotropic phase. For this reason, the order parameter $\boldsymbol{M}$ *can* describe the orientational ordering in the polar phase, but it *cannot* describe the orientational ordering in the nematic phase. The nematic phase clearly has some orientational order, but that order is not expressed in $\boldsymbol{M}$. Hence, we will call $\boldsymbol{M}$ the *polar order parameter,* and we will have to keep looking for a nematic order parameter.

The problem with using $\boldsymbol{M}$ to describe the nematic phase is that it is *odd* in the molecular orientation vectors, so that the contributions from molecules pointing up or down cancel each other. Clearly, we need a quantity that is *even* in the molecular orientation vectors. How about the average dot product $\langle\hat{\boldsymbol{\ell}}\cdot\hat{\boldsymbol{\ell}}\rangle$? No, that will not work because it is always 1, no matter what phase the system is in. How about the average cross product $\langle\hat{\boldsymbol{\ell}}\times\hat{\boldsymbol{\ell}}\rangle$? No, that will not work either because it is always 0. How about the average tensor product (or dyad)? This seems more promising. Let us tentatively define the tensor order parameter

$$\boldsymbol{T}=\langle\hat{\boldsymbol{\ell}}\otimes\hat{\boldsymbol{\ell}}\rangle=\langle\hat{\boldsymbol{\ell}}\hat{\boldsymbol{\ell}}\rangle, \tag{10.3}$$

or equivalently in terms of components

$$T_{\alpha\beta}=\langle\ell_\alpha\ell_\beta\rangle. \tag{10.4}$$

Can this tensor distinguish between the isotropic and nematic phases?

To answer this question, we first evaluate the tensor in the isotropic phase. The diagonal components are $T_{xx}=\langle\ell_x^2\rangle$, $T_{yy}=\langle\ell_y^2\rangle$, and $T_{zz}=\langle\ell_z^2\rangle$. These three components must be equal to each other because the phase is isotropic. Furthermore, these three components must add up to 1 because $\hat{\boldsymbol{\ell}}$ is a unit vector. Hence, each diagonal component must be $\frac{1}{3}$. The off-diagonal components are $T_{xy}=T_{yx}=\langle\ell_x\ell_y\rangle$, $T_{xz}=T_{zx}=\langle\ell_x\ell_z\rangle$, and $T_{yz}=T_{zy}=\langle\ell_y\ell_z\rangle$. Because the phase is isotropic, the three components ℓ_x, ℓ_y, and ℓ_z are equally likely to be positive or negative, and they fluctuate independently of each other. Hence, the off-diagonal components all average to 0. Putting these results together, the tensor becomes

$$T_{\alpha\beta}=\begin{pmatrix}\frac{1}{3}&0&0\\0&\frac{1}{3}&0\\0&0&\frac{1}{3}\end{pmatrix}=\frac{1}{3}\delta_{\alpha\beta} \tag{10.5}$$

in the isotropic phase.

Next, let us evaluate the tensor in a state with *perfect* nematic order along the z-axis. In this state, all of the molecules have the orientation $\hat{\boldsymbol{\ell}}=\pm\hat{z}$. As a result, the component $T_{zz}=1$ and all other components are zero. Hence, the tensor becomes

$$T_{\alpha\beta}=\begin{pmatrix}0&0&0\\0&0&0\\0&0&1\end{pmatrix}. \tag{10.6}$$

The results of Eqs. (10.5) and (10.6) show that the tensor $T_{\alpha\beta}$ *can* distinguish between the isotropic and nematic phases. However, they inspire us to make two small modifications in the order parameter. First, Eq. (10.5) shows that the isotropic average of $T_{\alpha\beta}$ is nonzero. It is more convenient to have an order parameter that is zero in the disordered phase. We can achieve that by subtracting off the isotropic average, to obtain $T'_{\alpha\beta} = T_{\alpha\beta} - \frac{1}{3}\delta_{\alpha\beta}$. Hence, $T'_{\alpha\beta} = 0$ in the isotropic phase, and

$$T'_{\alpha\beta} = \begin{pmatrix} -\frac{1}{3} & 0 & 0 \\ 0 & -\frac{1}{3} & 0 \\ 0 & 0 & \frac{2}{3} \end{pmatrix} \tag{10.7}$$

in the state with perfect nematic order along the z-axis. Second, Eq. (10.7) looks peculiar; it appears as if the perfect nematic state has only $\frac{2}{3}$ order. Just to make it look better, we will multiply this tensor by a factor of $\frac{3}{2}$. Hence, our final definition of the tensor order parameter for the nematic phase is $Q_{\alpha\beta} = \frac{3}{2}T'_{\alpha\beta}$, or

$$Q_{\alpha\beta} = \left\langle \frac{3}{2}\ell_\alpha \ell_\beta - \frac{1}{2}\delta_{\alpha\beta} \right\rangle. \tag{10.8}$$

We see that $Q_{\alpha\beta} = 0$ in the isotropic phase, and

$$Q_{\alpha\beta} = \begin{pmatrix} -\frac{1}{2} & 0 & 0 \\ 0 & -\frac{1}{2} & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{10.9}$$

in the state with perfect nematic order along the z-axis.

With this definition, we can see that the $Q_{\alpha\beta}$ tensor has two important mathematical properties. First, it is *symmetric,* meaning that $Q_{\alpha\beta} = Q_{\beta\alpha}$. Second, it is *traceless,* meaning that $\mathrm{Tr}(\boldsymbol{Q}) = Q_{\alpha\alpha} = 0$.

Now let us consider a state with only *partial* nematic order along the z-axis, as shown in the schematic illustration of Fig. 10.2c. Here, each molecule has its own unit vector $\hat{\boldsymbol{\ell}}$, which we can write in spherical coordinates as $\hat{\boldsymbol{\ell}} = (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta)$. The angle θ is the angle away from the z-axis, and it may be clustered around 0 and π. The angle ϕ is the azimuthal angle in the xy-plane, and it must be uniformly distributed between 0 and 2π. Hence, the azimuthal averages are $\langle\cos\phi\rangle = \langle\sin\phi\rangle = \langle\cos 2\phi\rangle = \langle\sin 2\phi\rangle = 0$ and $\langle\cos^2\phi\rangle = \langle\sin^2\phi\rangle = \frac{1}{2}$, independent of θ. Hence, the nematic order tensor has the component

$$Q_{zz} = \left\langle \frac{3}{2}\cos^2\theta - \frac{1}{2} \right\rangle = \langle P_2(\cos\theta)\rangle, \tag{10.10}$$

where $P_2(u) = \frac{3}{2}u^2 - \frac{1}{2}$ is the *second Legendre polynomial,* one of the special functions of mathematical physics, which you may have studied in classes on electrostatics or quantum mechanics. The average $\langle P_2(\cos\theta)\rangle$ is called the *scalar order parameter S* of the nematic phase,

$$S = \langle P_2(\cos\theta)\rangle = \left\langle \frac{3}{2}\cos^2\theta - \frac{1}{2}\right\rangle. \tag{10.11}$$

We will discuss the physical significance of S below. The other components of the nematic order tensor are

$$\begin{aligned}
Q_{xx} &= \left\langle \frac{3}{2}\sin^2\theta\cos^2\phi - \frac{1}{2}\right\rangle = \left\langle \frac{3}{4}\sin^2\theta - \frac{1}{2}\right\rangle = -\frac{S}{2},\\
Q_{yy} &= \left\langle \frac{3}{2}\sin^2\theta\sin^2\phi - \frac{1}{2}\right\rangle = \left\langle \frac{3}{4}\sin^2\theta - \frac{1}{2}\right\rangle = -\frac{S}{2},\\
Q_{xz} &= Q_{zx} = \left\langle \frac{3}{2}\cos\theta\sin\theta\cos\phi\right\rangle = 0,\\
Q_{yz} &= Q_{zy} = \left\langle \frac{3}{2}\cos\theta\sin\theta\sin\phi\right\rangle = 0,\\
Q_{xy} &= Q_{yx} = \left\langle \frac{3}{2}\sin^2\theta\cos\phi\sin\phi\right\rangle = \left\langle \frac{3}{4}\sin^2\theta\sin 2\phi\right\rangle = 0.
\end{aligned} \tag{10.12}$$

Putting these components together, the tensor representing partial nematic order along the z-axis is

$$Q_{\alpha\beta} = \begin{pmatrix} -\frac{S}{2} & 0 & 0\\ 0 & -\frac{S}{2} & 0\\ 0 & 0 & S\end{pmatrix}. \tag{10.13}$$

What if the nematic order is along the x-axis instead of the z-axis? In this case, the matrix component S must be in the xx position instead of the zz position, giving

$$Q_{\alpha\beta} = \begin{pmatrix} S & 0 & 0\\ 0 & -\frac{S}{2} & 0\\ 0 & 0 & -\frac{S}{2}\end{pmatrix}. \tag{10.14}$$

In this case, S is the average $\langle P_2(\cos\theta)\rangle$, where θ is the angle away from the x-axis. Similarly, if the nematic order is along the y-axis, then the order tensor must be

$$Q_{\alpha\beta} = \begin{pmatrix} -\frac{S}{2} & 0 & 0\\ 0 & S & 0\\ 0 & 0 & -\frac{S}{2}\end{pmatrix}, \tag{10.15}$$

where $S = \langle P_2(\cos\theta)\rangle$ and θ is the angle away from the y-axis.

What if the nematic order is along some arbitrary axis $\hat{\boldsymbol{n}}$, which is not $\hat{\boldsymbol{x}}$, $\hat{\boldsymbol{y}}$, or $\hat{\boldsymbol{z}}$? To answer this question, let us return to Eq. (10.13) for partial order along $\hat{\boldsymbol{z}}$. In this case, the order tensor can be rewritten as

$$
\begin{aligned}
Q_{\alpha\beta} =&S\left[\frac{3}{2}\begin{pmatrix}0&0&0\\0&0&0\\0&0&1\end{pmatrix}-\frac{1}{2}\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix}\right]\\
=&S\left[\frac{3}{2}z_\alpha z_\beta-\frac{1}{2}\delta_{\alpha\beta}\right],
\end{aligned}
\tag{10.16}
$$

where $z_\alpha = (0, 0, 1)$ are the components of $\hat{\boldsymbol{z}}$. Hence, if the nematic order is along an arbitrary axis $\hat{\boldsymbol{n}}$, the order tensor becomes

$$
Q_{\alpha\beta} = S\left[\frac{3}{2}n_\alpha n_\beta-\frac{1}{2}\delta_{\alpha\beta}\right],
\tag{10.17}
$$

where $S = \langle P_2(\cos\theta)\rangle$ and θ is the angle away from $\hat{\boldsymbol{n}}$.

Equation (10.17) is very important because it leads us to a physical interpretation of the tensor order parameter:

- The first factor of S, called the scalar order parameter, describes the *magnitude* of nematic order. It shows how well the molecules are aligned with each other. If $S = 0$, there is no alignment, and the system is in an isotropic phase. If $S = 1$, there is complete alignment, and the system is in a perfect nematic state. In general, for typical nematic liquid crystals, S is between 0 and 1, and the system has some partial nematic order.[2] This quantity is analogous to the magnitude $|M|$ in the Ising model, which shows how well the Ising spins are aligned with each other.
- The second factor of $[\frac{3}{2}n_\alpha n_\beta - \frac{1}{2}\delta_{\alpha\beta}]$ describes the *direction* of nematic order. The direction is encoded in this tensor in a complicated way, but it is in there! The vector $\hat{\boldsymbol{n}}$, with components n_α, is called the nematic *director*. It is a unit vector that represents the main axis of nematic order. It is analogous to the positive or negative sign of the Ising order parameter, which represents whether the net magnetic order is in the up or down direction.

  You should notice that $\hat{\boldsymbol{n}}$ is *always* a unit vector, no matter whether the nematic order is strong or weak. The magnitude of nematic order is not described by $\hat{\boldsymbol{n}}$ but by S. You should also notice that the order tensor $Q_{\alpha\beta}$ is even in $\hat{\boldsymbol{n}}$. As a result, $+\hat{\boldsymbol{n}}$ and $-\hat{\boldsymbol{n}}$ represent the same physical state; they are completely equivalent to each other.

  You might occasionally hear people say that "$\hat{\boldsymbol{n}}$ is the average orientation of the molecules." This is a very loose statement; people only say it because it is easy. You should not take it literally! When people say that, they really mean what I am saying here: $\hat{\boldsymbol{n}}$ is a unit vector representing the main axis of nematic order. (They really mean this statement even if they do not know that they mean it.)

[2] In principle, we should consider one more possibility: Because S is defined as $\langle P_2(\cos\theta)\rangle = \langle\frac{3}{2}\cos^2\theta - \frac{1}{2}\rangle$, it can theoretically be in the range $-\frac{1}{2} \leq S \leq 1$. If $S = 0$, the system is in an isotropic state. If $0 < S \leq 1$, the system is in a typical nematic state, with the molecular orientations attracted toward the axis $\hat{\boldsymbol{n}}$. If $-\frac{1}{2} \leq S < 0$, the system is in a peculiar *negative nematic* state, with the molecular orientations repelled away from the axis $\hat{\boldsymbol{n}}$. This negative nematic state is a theoretical possibility, but it does not normally occur in experiments.

Apart from the mathematical significance, Eq. (10.17) provides a nice way to think about liquid-crystal science:

- Part of liquid-crystal science involves controlling the magnitude of order. This magnitude is mainly controlled by temperature in a pure liquid crystal, or by concentration in a colloidal suspension. In this chapter, it will be the subject of Sects. 10.3–10.5. In general, I would estimate that it is about 10 % of liquid-crystal science.
- The rest of liquid-crystal science involves controlling the direction of order. This direction is controlled by electric and magnetic fields, surface anchoring, shear flow, and everything else that one might do to a liquid crystal in the laboratory. In this chapter, it will be the subject of Sects. 10.6–10.9. In general, I would estimate that it is about 90 % of liquid-crystal science.

If we know the tensor components $Q_{\alpha\beta}$, how can we determine S and $\hat{\boldsymbol{n}}$? This question brings us back to Sect. 9.6 about magnitudes and directions of tensors. Like any symmetric tensor, $Q_{\alpha\beta}$ has magnitudes given by its eigenvalues, and directions given by its eigenvectors. Hence, we should diagonalize $Q_{\alpha\beta}$ to find the eigenvalues and eigenvectors. The three eigenvalues are S, $-S/2$, and $-S/2$. (Because $Q_{\alpha\beta}$ is traceless, these three eigenvalues add up to zero.) The eigenvalue S is associated with the eigenvector $\hat{\boldsymbol{n}}$. The eigenvalue $-S/2$ is associated with any eigenvector perpendicular to $\hat{\boldsymbol{n}}$. (Because there are two degenerate eigenvalues of $-S/2$, there is a freedom to rotate the eigenvectors in the plane perpendicular to $\hat{\boldsymbol{n}}$.)

Problem: Show explicitly that $\boldsymbol{Q} \cdot \hat{\boldsymbol{n}} = S\hat{\boldsymbol{n}}$. This equation proves that $\hat{\boldsymbol{n}}$ is an eigenvector of $\boldsymbol{Q}$ associated with eigenvalue S.

Solution: Equation (10.17) gives

$$\begin{aligned}(\boldsymbol{Q} \cdot \hat{\boldsymbol{n}})_\alpha = Q_{\alpha\beta} n_\beta = S\left[\frac{3}{2} n_\alpha n_\beta - \frac{1}{2}\delta_{\alpha\beta}\right] n_\beta = S\left[\frac{3}{2} n_\alpha n_\beta n_\beta - \frac{1}{2}\delta_{\alpha\beta} n_\beta\right] \\ = S\left[\frac{3}{2} n_\alpha - \frac{1}{2} n_\alpha\right] = S n_\alpha. \qquad (10.18)\end{aligned}$$

Problem: Assuming that $\hat{\boldsymbol{c}}$ is perpendicular to $\hat{\boldsymbol{n}}$, show explicitly that $\boldsymbol{Q} \cdot \hat{\boldsymbol{c}} = -(S/2)\hat{\boldsymbol{c}}$. This equation proves that $\hat{\boldsymbol{c}}$ is an eigenvector of $Q_{\alpha\beta}$ associated with eigenvalue $-S/2$.

Solution: Again using Eq. (10.17),

$$\begin{aligned}(\boldsymbol{Q} \cdot \hat{\boldsymbol{c}})_\alpha = Q_{\alpha\beta} c_\beta = S\left[\frac{3}{2} n_\alpha n_\beta - \frac{1}{2}\delta_{\alpha\beta}\right] c_\beta = S\left[\frac{3}{2} n_\alpha n_\beta c_\beta - \frac{1}{2}\delta_{\alpha\beta} c_\beta\right] \\ = S\left[0 - \frac{1}{2} c_\alpha\right] = -\left(\frac{S}{2}\right) c_\alpha. \qquad (10.19)\end{aligned}$$

As a final point in this section, let us count the degrees of freedom associated with the tensor order parameter. In 3D, an arbitrary second-rank tensor has nine degrees of freedom. Because $Q_{\alpha\beta}$ is symmetric, three degrees of freedom are eliminated. Because $Q_{\alpha\beta}$ is traceless, one more is eliminated. Hence, the nematic order tensor has five degrees of freedom.

Of these five degrees of freedom, one is associated with the scalar order parameter S. Two more are associated with the director $\hat{\boldsymbol{n}}$—for example, they could be the angles θ and ϕ for this unit vector. Hence, we understand three degrees of freedom.

What are the other two degrees of freedom? Well, the tensor $Q_{\alpha\beta}$ is more general than the nematic phase that we have discussed. We have discussed the usual type of nematic phase, which is *uniaxial,* meaning that it has only one special axis $\hat{\boldsymbol{n}}$. All of the directions perpendicular to $\hat{\boldsymbol{n}}$ are equivalent to each other, i.e., there is perfect rotational symmetry about $\hat{\boldsymbol{n}}$. By comparison, an alternative theoretical possibility would be a *biaxial* nematic phase, in which the rotational symmetry about $\hat{\boldsymbol{n}}$ is broken. In this case, the phase would have three special axes, all perpendicular to each other. Each of these axes would be associated with a distinct eigenvalue of $Q_{\alpha\beta}$. Because the tensor is still traceless, these eigenvalues would have to add up to zero, and hence they could be written as S, $-(S+P)/2$, and $-(S-P)/2$. Here, P can be regarded as a biaxial order parameter. The biaxial nematic phase would have one degree of freedom associated with P, and one more degree of freedom associated with the orientation of biaxial order in the plane perpendicular to $\hat{\boldsymbol{n}}$. This accounts for the extra two degrees of freedom. Although the biaxial nematic phase is a theoretical possibility, there is still some controversy about whether it occurs in experiments; this is an active area of research.

10.3 Landau-de Gennes Theory

As noted above, one important part of liquid-crystal science is to understand what determines the magnitude S of nematic order. In particular, we want to understand whether S is zero (in the isotropic phase) or nonzero (in the nematic phase). In other words, we want to develop a model of the isotropic-nematic transition.

You will recognize that this problem is quite analogous to the Ising model, where determined the magnitude $|M|$ of magnetic order, and we developed a model for the transition from $|M| = 0$ to $|M| \neq 0$. Hence, we can use the Ising model as an example to help us understand the isotropic-nematic transition.

When we discussed the Ising model, we used two types of theoretical approaches. In Chap. 2, we began with a microscopic model of interacting spins, and used mean-field theory to calculate the statistical ordering of the spins. By comparison, in Chap. 4, we used the purely macroscopic approach of Landau theory to construct the free energy in terms of the Ising order parameter, and minimized this free energy. Each of these approaches has its own advantages and disadvantages. The microscopic theory makes very specific predictions, but it applies only to one particular model.

The macroscopic theory makes some predictions for a broad class of problems, but it does not make very specific predictions for any of them.

Both of these theoretical approaches can be applied to the isotropic-nematic transition. There are two well-known microscopic theories, Maier-Saupe theory and Onsager theory, which we will discuss in the following sections. However, we will begin with the macroscopic approach of Landau theory. When applied to the isotropic-nematic transition, this theoretical approach is often called Landau-de Gennes theory, after the great French physicist Pierre-Gilles de Gennes, who developed so much of modern liquid-crystal science.

As you recall from Chap. 4, the basic concept of Landau theory is to construct the free energy as a function of the order parameter, in the most general possible way that is permitted by symmetry. The free energy is always a scalar, meaning that it is invariant under rotation. For the isotropic-nematic transition, the order parameter is the second-rank tensor $Q_{\alpha\beta}$. Hence, we must find ways to construct a scalar out of a second-rank tensor. In this construction, we make two important assumptions. First, we assume the free energy is a smooth (analytic) function of the order parameter; we want to see how a smooth *input* to the theory leads to a phase transition, i.e., a non-smooth *output* from the theory. Because the free energy is a smooth function, it can be represented as a power series in the order parameter. Second, we assume that the order parameter is small, so that the power series only needs a small number of terms.

What terms can go into our power series for the free energy? At zeroth order in $Q_{\alpha\beta}$, there can be a constant term, which we can just call F_0. At first order, the only way to make a scalar out of a second-rank tensor is to take the trace $Q_{\alpha\alpha}$. However, we defined the order parameter in the previous section so that it is traceless, $Q_{\alpha\alpha} = 0$. Hence, there can be no first-order term in the power series. (An exception is if there is some symmetry-breaking field in the problem, as discussed below.) At second order, we can make a scalar out of two powers of the tensor as $Q_{\alpha\beta}Q_{\alpha\beta}$, implicitly summed over α and β. At third order, we can make a scalar out of three powers of the tensor as $Q_{\alpha\beta}Q_{\beta\gamma}Q_{\gamma\alpha}$. At fourth order, we can make a scalar out of four powers of the tensor as $(Q_{\alpha\beta}Q_{\alpha\beta})^2$ or as $Q_{\alpha\beta}Q_{\beta\gamma}Q_{\gamma\delta}Q_{\delta\alpha}$. Putting these terms together, we can write the free energy density (per unit volume) as the power series

$$\begin{aligned} f = \frac{F}{V} = &f_0 + \frac{1}{2}AQ_{\alpha\beta}Q_{\alpha\beta} + \frac{1}{3}BQ_{\alpha\beta}Q_{\beta\gamma}Q_{\gamma\alpha} + \frac{1}{3}C_1(Q_{\alpha\beta}Q_{\alpha\beta})^2 \\ &+ \frac{1}{3}C_2Q_{\alpha\beta}Q_{\beta\gamma}Q_{\gamma\delta}Q_{\delta\alpha} + \cdots, \end{aligned} \tag{10.20}$$

where all the coefficients are arbitrary.

Now let us re-express the free energy in terms of the magnitude S and director $\hat{\boldsymbol{n}}$ of nematic order. Using Eq. (10.17), we obtain

$$\begin{aligned}
Q_{\alpha\beta}Q_{\alpha\beta} &= S\left[\frac{3}{2}n_\alpha n_\beta - \frac{1}{2}\delta_{\alpha\beta}\right] S\left[\frac{3}{2}n_\alpha n_\beta - \frac{1}{2}\delta_{\alpha\beta}\right] \\
&= S^2\left[\frac{9}{4}n_\alpha n_\beta n_\alpha n_\beta - 2\frac{3}{4}n_\alpha n_\beta \delta_{\alpha\beta} + \frac{1}{4}\delta_{\alpha\beta}\delta_{\alpha\beta}\right] \qquad (10.21) \\
&= S^2\left[\frac{9}{4}n_\alpha n_\alpha n_\beta n_\beta - \frac{3}{2}n_\alpha n_\alpha + \frac{1}{4}\delta_{\alpha\alpha}\right] = S^2\left[\frac{9}{4} - \frac{3}{2} + \frac{3}{4}\right] = \frac{3}{2}S^2.
\end{aligned}$$

Similarly,

$$\begin{aligned}
Q_{\alpha\beta}Q_{\beta\gamma}Q_{\gamma\alpha} &= \frac{3}{4}S^3, \\
(Q_{\alpha\beta}Q_{\alpha\beta})^2 &= \frac{9}{4}S^4, \qquad (10.22) \\
Q_{\alpha\beta}Q_{\beta\gamma}Q_{\gamma\delta}Q_{\delta\alpha} &= \frac{9}{8}S^4.
\end{aligned}$$

Hence, the power series becomes

$$f = f_0 + \frac{1}{2}aS^2 + \frac{1}{3}bS^3 + \frac{1}{4}cS^4 + \cdots, \qquad (10.23)$$

where $a = \frac{3}{2}A$, $b = \frac{3}{4}B$, and $c = \frac{9}{4}C_1 + \frac{9}{8}C_2$.

How does the free energy depend on the director $\hat{\boldsymbol{n}}$? It does not! Because there is no symmetry-breaking field applied to the system, the nematic order is equally likely to form in any direction. The system will randomly select an orientation for nematic order. This is an example of *spontaneous symmetry breaking.*

The Landau-de Gennes free energy is expressed in terms of the unknown coefficients a, b, and c. We suppose that these coefficients are all smooth functions of temperature, and consider what happens to the shape of the free energy plot as these coefficients change. This behavior depends on whether b and c are positive or negative. If we are going to truncate the series at fourth order in S, we must have $c > 0$. (If c were negative, then the free energy would go to $-\infty$ as S became large, and hence the system would be unstable.) We do not know in advance whether b is positive or negative. For now, let us assume that $b < 0$, which will turn out to be the normal physical case. (We will discuss $b > 0$ briefly below.)

Figure 10.3 shows a series of plots of the free energy as the parameter a decreases, presumably because of decreasing temperature, for fixed negative b and positive c. At high temperature, for large positive a, the free energy has the form in Fig. 10.3a. Here, the free energy has only one minimum, which is at $S = 0$, corresponding to the isotropic phase. As the temperature decreases and a decreases, the free energy changes to the form shown in Fig. 10.3b. Now it has two minima: the stable isotropic minimum at $S = 0$ and a metastable nematic minimum at $S > 0$. As a continues to decrease, the nematic minimum becomes deeper. At a certain value of a, shown in Fig. 10.3c, the isotropic and nematic minima become equally deep. At this point,

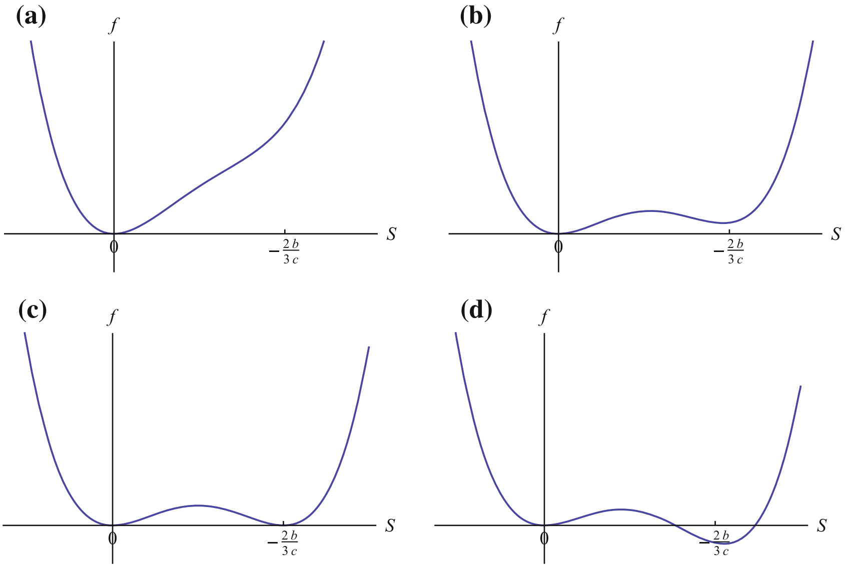


Fig. 10.3 Plots of the Landau-de Gennes free energy density f as a function of the nematic order parameter S. The parameters b and c are held fixed, with $b < 0$ and $c > 0$, and the parameter a is varied. **a** Far above the isotropic-nematic transition. **b** Slightly above the isotropic-nematic transition. **c** At the the isotropic-nematic transition. **d** Below the isotropic-nematic transition (Interactive version at http://www.springer.com/cda/content/document/cda_downloaddocument/Selinger+Interactive+Figures.zip?SGWID=0-0-45-1509169-p177545420.)

we see a *first-order transition* from the isotropic phase to the nematic phase. Below that transition, the free energy takes the form shown in Fig. 10.3d. It has a stable minimum in the nematic phase ($S > 0$), and a metastable minimum in the isotropic phase ($S = 0$).

To find the nematic order parameter that minimizes the free energy, we solve the equation

$$\frac{\partial f}{\partial S} = 0. \tag{10.24}$$

The solutions are

$$S = 0 \text{ or } S = \frac{-b \pm \sqrt{b^2 - 4ac}}{2c}. \tag{10.25}$$

There are either one or three real solutions, depending on whether the quantity inside the square root is negative or positive. Figure 10.3a shows a case with one real solution at $S = 0$. Figure 10.3b–d shows cases with three real solutions: the isotropic minimum at $S = 0$, the maximum between the isotropic and nematic states (solution with the negative square root), and the nematic minimum (solution with the positive square root).

To determine which minimum is lowest, we evaluate the free energy at each minimum and compare them. This calculation shows that the first-order isotropic-nematic transition occurs at

$$a_{IN} = \frac{2b^2}{9c}, \tag{10.26}$$

and the order parameter on the nematic side of the transition is

$$S = -\frac{2b}{3c}. \tag{10.27}$$

This calculation shows the value of a at the transition. Of course, it would be much more useful to know the temperature at the transition. In Landau theory, as discussed in Chap. 4, people normally assume that a varies linearly with temperature, and express it as

$$a = a'(T - T_0). \tag{10.28}$$

Hence, Eq. (10.26) for the isotropic-nematic transition becomes

$$T_{IN} = T_0 + \frac{2b^2}{9a'c}. \tag{10.29}$$

You should notice that the first-order transition occurs at a temperature where a has a certain positive value, unlike the second-order transitions discussed in Chap. 4, which occur at a temperature where $a = 0$.

Figure 10.4 shows a plot of the prediction for the order parameter as a function of temperature. For $T > T_{IN}$, the system is in the isotropic phase with $S = 0$. At $T = T_{IN}$, the order parameter jumps discontinuously from 0 to $-2b/3c$. As T continues to decrease below T_{IN}, the order parameter continues to increase, following Eq. (10.25) (solution with the positive square root). This is the general trend observed in experiments on the isotropic-nematic transition.

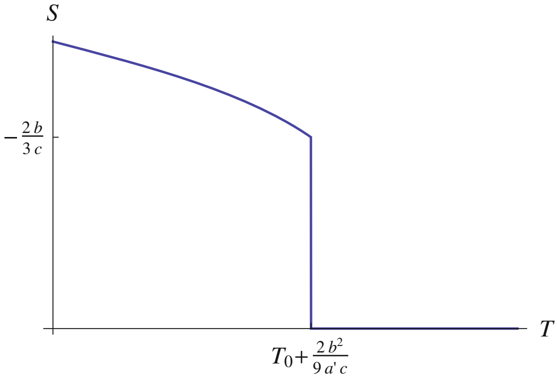


Fig. 10.4 Prediction of Landau-de Gennes theory for the nematic order parameter S as a function of temperature T, showing the first-order isotropic-nematic transition

So far, the calculation has assumed that the coefficient b is negative. By comparison, if b is positive, then the behavior will be very similar but the nematic minimum will occur for $S < 0$. This is the negative nematic state mentioned briefly in the previous section. It is a theoretical possibility but it generally does not occur in experiments, and hence we will not consider it further.

To summarize the theory so far, we have learned two important points. First, the isotropic-nematic transition is a first-order transition, with a discontinuity in the nematic order parameter. Second, the order parameter has the general temperature dependence shown in Fig. 10.4. Those are significant accomplishments for a theory based only on general assumptions about symmetry and smooth functions! However, the theory has some limitations. One limitation is that the predictions are expressed in terms of the arbitrary parameters a, b, and c, and we do not actually know these parameters. Another limitation is that this prediction for S does not reach a maximum at 1; rather, it continues increasing without limit as the temperature decreases. This part of the prediction is unphysical; the physical order parameter cannot be any larger than 1. As discussed in Chap. 4, this problem arises because Landau-de Gennes theory is a power series about the disordered phase, and hence is not aware of the maximum possible degree of order. In order to go beyond these limitations, we will need to go to a more microscopic theory, as discussed in the following two sections.

Before going on, let us consider one more point that we can learn from Landau-de Gennes theory. Suppose that we apply a symmetry-breaking field to the material. To be specific, let us consider an magnetic field $\boldsymbol{H}$, although it might also be an electric field. In this case, the magnetic field will couple to the nematic order tensor $Q_{\alpha\beta}$, leading to additional terms in the free energy density. The most important new term is the first-order term

$$E_{\text{coupling}} = -\frac{\Delta\chi_{\text{max}}}{3\mu_0} H_\alpha H_\beta Q_{\alpha\beta}. \tag{10.30}$$

Here, the coefficient $\Delta\chi_{\text{max}}$ is the maximum diamagnetic anisotropy of the fully aligned ($S = 1$) liquid crystal. (We will discuss the diamagnetic anisotropy further in Sect. 10.7.) This term is a scalar, which is constructed from a combination of the magnetic field and the order tensor, summed over α and β. If we express $Q_{\alpha\beta}$ in terms of S and $\hat{\boldsymbol{n}}$, this term reduces to

$$E_{\text{coupling}} = -\frac{\Delta\chi_{\text{max}}}{3\mu_0} S \left[\frac{3}{2}(\boldsymbol{H}\cdot\hat{\boldsymbol{n}})^2 - \frac{1}{2}|\boldsymbol{H}|^2\right]. \tag{10.31}$$

Unlike the other free energy terms discussed in this section, this term depends on $\hat{\boldsymbol{n}}$ as well as S. Let us assume that $\Delta\chi_{\text{max}}$ and S are both positive, so this term favors alignment of $\hat{\boldsymbol{n}}$ with $\boldsymbol{H}$. (This is the case of *positive* diamagnetic anisotropy; the opposite case is also possible.) If $\hat{\boldsymbol{n}}$ is aligned with $\boldsymbol{H}$, the coupling term reduces to

$$E_{\text{coupling}} = -\frac{\Delta\chi_{\text{max}}}{3\mu_0}|\boldsymbol{H}|^2 S. \tag{10.32}$$

Hence, the full free energy density becomes

$$f = f_0 - \frac{\Delta\chi_{\max}}{3\mu_0}|\boldsymbol{H}|^2 S + \frac{1}{2}aS^2 + \frac{1}{3}bS^3 + \frac{1}{4}cS^4 + \cdots . \tag{10.33}$$

Minimizing the free energy over S gives

$$\frac{\partial f}{\partial S} = -\frac{\Delta\chi_{\max}}{3\mu_0}|\boldsymbol{H}|^2 + aS + bS^2 + cS^3 = 0. \tag{10.34}$$

In general, this equation is difficult to solve. However, it simplifies greatly in the limit of low field and high temperature (large positive a). In this case, we expect the order parameter S to be very small. In this case, we can neglect the b and c terms in the free energy series, and consider only the $\Delta\chi_{\max}$ and a terms. The minimum then occurs at

$$S = \frac{\Delta\chi_{\max}|\boldsymbol{H}|^2}{3\mu_0 a} = \frac{\Delta\chi_{\max}|\boldsymbol{H}|^2}{3\mu_0 a'(T - T_0)}. \tag{10.35}$$

This equation shows us that a magnetic field induces a small amount of nematic order for $T > T_{IN}$. This induced order is called the *Cotton–Mouton effect.* The induced order is proportional to $|\boldsymbol{H}|^2$, not proportional to $\boldsymbol{H}$; this quadratic dependence is a consequence of nematic order being a second-rank tensor. The induced order also depends on temperature; it is small at high T, and it increases as T approaches T_{IN}. Similar order is also induced by an electric field, in which case it is called the *Kerr effect.*

You will notice that field-induced order in the isotropic phase of liquid crystals is analogous to field-induced order in the paramagnetic phase of the Ising model. Based on this analogy, people sometimes refer to the isotropic phase under a field as a *paranematic* phase. This term is useful because the phase is not exactly isotropic; the field is breaking the rotational symmetry.

What happens if the field becomes larger and the temperature becomes lower, so that we cannot use the approximation of Eq. (10.35)? In this case, we can minimize the free energy of Eq. (10.33) numerically. Figure 10.5 shows a plot of the numerical solution for S as a function of T, for several values of the magnetic field (increasing from left to right). For zero field, we see the first-order transition from isotropic to nematic. When a small field is applied, it induces a small amount of order in the isotropic phase (which is now paranematic), and it slightly increases the order in the nematic phase. The system now has a first-order transition from the slightly ordered paranematic phase to the highly ordered nematic phase. The transition temperature is slightly higher than T_{IN}, and the discontinuity in the order parameter is slightly smaller than the original isotropic-nematic discontinuity. As the field increases, the transition temperature becomes even higher, and the discontinuity at the transition becomes even smaller. When the applied field becomes big enough, the discontinuity at the transition goes to zero. Instead of a discontinuity, there is just a point of infinite

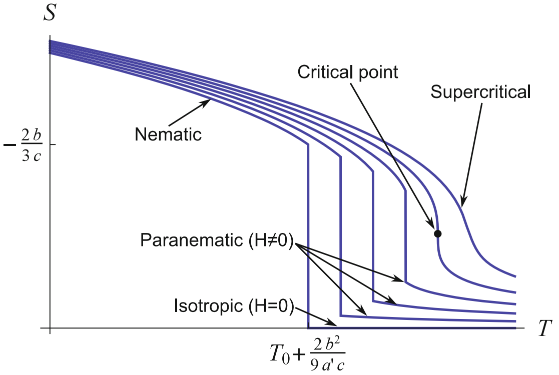


Fig. 10.5 Prediction of Landau-de Gennes theory for the nematic order parameter S as a function of temperature T, for several values of the magnetic field (increasing from *left* to *right*), showing the field-induced order and critical point

slope on the plot. This is a critical point! Beyond this critical point, the distinction between paranematic and nematic vanishes, and there is only a single supercritical state.

We can see that the nematic-paranematic critical point is analogous to the liquid-gas critical point studied in Chap. 3. These critical points occur for the same fundamental reason: The distinction between the liquid and gas phases, or between the nematic and paranematic phases, is not a symmetry difference. Rather, it is just a quantitative distinction in the magnitude of an order parameter (density for liquid and gas, S for nematic and paranematic). When there is a first-order transition with a quantitative discontinuity in some order parameter, this quantitative discontinuity can become larger or smaller, depending on applied fields that couple to the order parameter. In particular, the discontinuity can be driven to zero, leading to a critical point.

It is remarkable that Landau-de Gennes theory can provide this information about field-induced order and the field-induced critical point, based only on macroscopic considerations. However, we still might want to determine how the isotropic-nematic transition is related to microscopic interactions. For this, we will need to move on to a different type of theory, in the following two sections.

10.4 Maier-Saupe Theory

Maier-Saupe theory is a microscopic model that begins with an assumption about interactions between molecules, and then uses the mean-field approximation to calculate the energy, entropy, and free energy. By minimizing the free energy, it determines the isotropic-nematic transition temperature, as well as the order parameter in the nematic phase. In this respect, it is analogous to mean-field theory for the Ising model, which we discussed in Chap. 2.

Before beginning this calculation, we must deal with one annoying bit of notation. Everyone uses the letter S to represent the entropy, and everyone also uses the letter

S to represent the nematic order parameter. In this section, we will need to discuss both of these quantities, and I do not want you to get confused between them! For this reason, I will use S_{entropy} to represent the entropy, and just S to represent the nematic order parameter.

Now let us begin with an assumption about the interactions between molecules. Suppose that a system contains N molecules, and each molecule interacts with q neighboring molecules. Hence, the system contains $\frac{1}{2}Nq$ interacting pairs of molecules. For each pair, the interaction potential depends on the relative orientation of the two molecules in the pair. Suppose one molecule in the pair has the orientation $\hat{\boldsymbol{\ell}}$, and the other has the orientation $\hat{\boldsymbol{m}}$. In this case, the relative orientation angle γ is given by $\cos\gamma = \hat{\boldsymbol{\ell}} \cdot \hat{\boldsymbol{m}}$. Maier-Saupe theory assumes that the interaction potential is

$$V_{\text{int}} = -JP_2(\cos\gamma) = -J\left[\frac{3}{2}\cos^2\gamma - \frac{1}{2}\right] = -J\left[\frac{3}{2}\left(\hat{\boldsymbol{\ell}} \cdot \hat{\boldsymbol{m}}\right)^2 - \frac{1}{2}\right], \quad (10.36)$$

where J is some positive constant. You should notice that this potential is minimum when either $\hat{\boldsymbol{\ell}} = \hat{\boldsymbol{m}}$ or $\hat{\boldsymbol{\ell}} = -\hat{\boldsymbol{m}}$; it is maximum when $\hat{\boldsymbol{\ell}}$ and $\hat{\boldsymbol{m}}$ are perpendicular to each other. Hence, it favors nematic alignment of molecules along the same axis.

There are two ways to justify this assumption about the interaction potential. First, one can do a quantum-mechanical calculation of the interaction between a fluctuating electric dipole on one molecule and the induced dipole on the other molecule. This calculation gives an orientation-dependent interaction potential of the Maier-Saupe form. (It is an anisotropic version of the van der Waals interaction, which you may have seen in classes on colloid science.) Alternatively, one can say that there is some arbitrary orientation-dependent interaction between the molecules, and expand it as a series of spherical harmonics. The Maier-Saupe form is the leading term in this series. In either case, let us assume that this interaction is correct and proceed with the calculation.

It is impossible to work out the exact statistical mechanics of many molecules interacting through the Maier-Saupe potential. However, we can use mean-field theory to obtain an approximation to the free energy, as we did for the Ising model. In mean-field theory, we need to determine the energy and the entropy. The expectation value of the energy is the number of interacting pairs times the expectation value of the pair interaction

$$\langle E\rangle = -\frac{1}{2}NqJ\,\langle P_2(\cos\gamma)\rangle = -\frac{1}{2}NqJ\left[\frac{3}{2}\left\langle\left(\hat{\boldsymbol{\ell}} \cdot \hat{\boldsymbol{m}}\right)^2\right\rangle - \frac{1}{2}\right]. \quad (10.37)$$

To simplify this expression, we write the vectors in tensor component notation

$$\langle E\rangle = -\frac{1}{2}NqJ\left[\frac{3}{2}\left\langle\ell_\alpha m_\alpha \ell_\beta m_\beta\right\rangle - \frac{1}{2}\right] = -\frac{1}{2}NqJ\left[\frac{3}{2}\left\langle\ell_\alpha \ell_\beta m_\alpha m_\beta\right\rangle - \frac{1}{2}\right]. \quad (10.38)$$

We now make the mean-field approximation: We assume that the molecules are fluctuating independently and neglect correlations between their orientations. Hence, we assume that

$$\left\langle \ell_\alpha \ell_\beta m_\alpha m_\beta \right\rangle \approx \left\langle \ell_\alpha \ell_\beta \right\rangle \left\langle m_\alpha m_\beta \right\rangle . \tag{10.39}$$

This approximation is certainly not exactly true, but it will allow us to proceed. In this equation, the expectation value of the tensor product is related to the tensor order parameter of the nematic phase

$$Q_{\alpha\beta} = \frac{3}{2} \left\langle \ell_\alpha \ell_\beta \right\rangle - \frac{1}{2}\delta_{\alpha\beta} = \frac{3}{2} \left\langle m_\alpha m_\beta \right\rangle - \frac{1}{2}\delta_{\alpha\beta}, \tag{10.40}$$

and hence

$$\left\langle \ell_\alpha \ell_\beta \right\rangle = \left\langle m_\alpha m_\beta \right\rangle = \frac{2}{3} Q_{\alpha\beta} + \frac{1}{3}\delta_{\alpha\beta}. \tag{10.41}$$

Thus, the expectation value of the energy becomes

$$\begin{aligned} \langle E \rangle &= -\frac{1}{2} NqJ \left[\frac{3}{2} \left(\frac{2}{3} Q_{\alpha\beta} + \frac{1}{3}\delta_{\alpha\beta} \right) \left(\frac{2}{3} Q_{\alpha\beta} + \frac{1}{3}\delta_{\alpha\beta} \right) - \frac{1}{2} \right] \\ &= -\frac{1}{2} NqJ \left[\frac{3}{2} \left(\frac{4}{9} Q_{\alpha\beta} Q_{\alpha\beta} + \frac{1}{3} \right) - \frac{1}{2} \right] = -\frac{1}{3} NqJ Q_{\alpha\beta} Q_{\alpha\beta} \\ &= -\frac{1}{2} NqJS^2. \end{aligned} \tag{10.42}$$

This energy favors the maximum possible value of nematic order.

Next we consider the entropy, working by analogy with the Ising model. In the Ising model, the distribution of spin directions is just given by two probabilities, $p_\uparrow$ and $p_\downarrow$, which are normalized so that $p_\uparrow + p_\downarrow = 1$. In terms of these probabilities, the entropic contribution to the free energy was found to be

$$- T S_{\text{entropy}} = N k_B T \left(p_\uparrow \log p_\uparrow + p_\downarrow \log p_\downarrow \right) . \tag{10.43}$$

For a liquid crystal, the analogous concept is the *distribution function* for molecular orientations, which can be written as $\rho(\hat{\boldsymbol{\ell}})$ or $\rho(\theta, \phi)$, where θ and ϕ represent the molecular orientation in spherical coordinates. This distribution function must be normalized so that it integrates to 1:

$$\int_0^\pi \sin\theta d\theta \int_0^{2\pi} d\phi \rho(\theta, \phi) = 1. \tag{10.44}$$

Note that we are using the appropriate measure of integration in spherical coordinates. In terms of this orientational distribution function, the entropic contribution to the free energy becomes

$$-TS_{\text{entropy}} = Nk_BT\int_0^{\pi}\sin\theta d\theta\int_0^{2\pi}d\phi\rho(\theta,\phi)\log\rho(\theta,\phi). \tag{10.45}$$

When we combine the energy and entropy, we obtain the free energy for Maier-Saupe theory

$$\begin{aligned} F &= \langle E\rangle - TS_{\text{entropy}} \qquad (10.46)\\ &= -\frac{1}{2}NqJS^2 + Nk_BT\int_0^{\pi}\sin\theta d\theta\int_0^{2\pi}d\phi\rho(\theta,\phi)\log\rho(\theta,\phi). \end{aligned}$$

In this expression, the nematic order parameter in the first term is determined by the orientational distribution function as

$$S = \langle P_2(\cos\theta)\rangle = \int_0^{\pi}\sin\theta d\theta\int_0^{2\pi}d\phi P_2(\cos\theta)\rho(\theta,\phi). \tag{10.47}$$

The question is now: What orientational distribution function gives the minimum free energy? One approach to answer that question relies on some physical reasoning, as follows. (If you do not believe that physical reasoning, you can read the more mathematical argument at the end of this section.)

For a physical approach, consider a single molecule surrounded by its neighbors. This test molecule experiences an effective potential, which arises from its interactions with all of the neighbors. If the neighbors have some alignment along the z-axis, then the effective potential will tend to align the test molecule along the z-axis also. By comparison, if the neighbors have an isotropic distribution of orientations, then the effective potential will not align the test molecule in any particular direction. In general, we expect that the effective potential must have the form

$$V_{\text{eff}}(\theta) = -Uk_BTP_2(\cos\theta), \tag{10.48}$$

where Uk_BT is the characteristic strength of the potential, or U is the potential strength divided by k_BT. We do not know U yet, but we will calculate it soon!

With this effective potential, the orientational distribution function becomes the Boltzmann distribution

$$\rho(\theta,\phi) = \frac{1}{D(U)}e^{-V_{\text{eff}}(\theta)/k_BT} = \frac{1}{D(U)}e^{UP_2(\cos\theta)}, \tag{10.49}$$

where the denominator is fixed by the normalization condition,

$$D(U) = \int_0^{\pi}\sin\theta d\theta\int_0^{2\pi}d\phi e^{UP_2(\cos\theta)} = 2\pi\int_0^{\pi}\sin\theta d\theta e^{UP_2(\cos\theta)}. \tag{10.50}$$

We can now express the nematic order parameter S in terms of U,

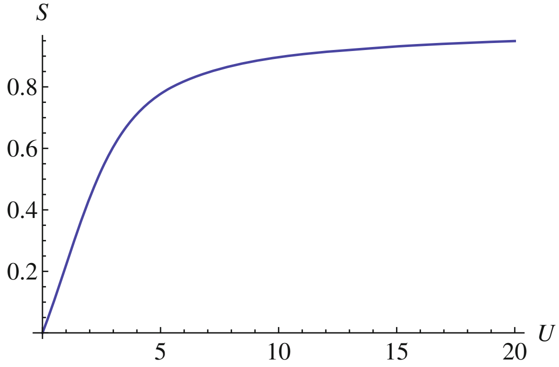


Fig. 10.6 Nematic order parameter S as a function of the variable U in the effective potential

$$S(U) = \frac{\int_0^\pi \sin\theta d\theta P_2(\cos\theta) e^{U P_2(\cos\theta)}}{\int_0^\pi \sin\theta d\theta e^{U P_2(\cos\theta)}}. \tag{10.51}$$

Unfortunately, there is no simple analytic expression for this ratio of integrals. However, we can certainly do the integrals numerically and plot the result for S as a function of U, as shown in Fig. 10.6. This plot basically shows that U is an alternative version of the nematic order parameter, which carries the same information in a different form. For weak nematic order, S and U are linearly proportional to each other. For strong nematic order, S saturates at 1, while U goes to ∞.

As a next step, we can express the Maier-Saupe free energy as a function of U. From Eq. (10.46), we obtain

$$\begin{aligned} F(U) =& -\frac{1}{2} N q J \left[S(U)\right]^2 \\ &+ N k_B T \int_0^\pi \sin\theta d\theta \int_0^{2\pi} d\phi \rho(\theta,\phi) \left[U P_2(\cos\theta) - \log D(U)\right] \\ =& -\frac{1}{2} N q J \left[S(U)\right]^2 + N k_B T \left[U S(U) - \log D(U)\right]. \end{aligned} \tag{10.52}$$

Normalizing by $N k_B T$ gives

$$\frac{F(U)}{N k_B T} = -\frac{Jq}{2 k_B T} \left[S(U)\right]^2 + U S(U) - \log D(U). \tag{10.53}$$

Hence, our strategy should be to plot the free energy of Eq. (10.53) as a function of U, and search for the minimum.

If we differentiate the free energy with respect to U, we obtain

$$\begin{aligned} \frac{\partial}{\partial U}\left(\frac{F(U)}{N k_B T}\right) =& -\frac{Jq}{k_B T} S(U) \frac{\partial S}{\partial U} + U \frac{\partial S}{\partial U} + S(U) - \frac{1}{D(U)} \frac{\partial D}{\partial U} \\ =& \frac{\partial S}{\partial U}\left[U - \frac{Jq}{k_B T} S(U)\right]. \end{aligned} \tag{10.54}$$

Hence, our condition for a minimum is

$$U = \frac{Jq}{k_B T} S(U). \tag{10.55}$$

This equation can be regarded as a self-consistency equation for U: The nematic order parameter S of a test molecule is determined by the effective potential strength U and, at the free energy minimum, the effective potential strength U is proportional to the nematic order parameter S of the neighboring molecules. One self-consistent solution is the isotropic state ($U = 0$ and $S = 0$), and there may also be a self-consistent nematic state. The equation can be solved numerically to find all the self-consistent states. However, it is more instructive to just plot the free energy and look for the minima.

Figure 10.7 shows a series of plots of the free energy as the scaled temperature $k_B T/Jq$ decreases. At high temperature, in Fig. 10.7a, the free energy has only one minimum, which is at $U = 0$, and hence $S = 0$, corresponding to the isotropic phase.

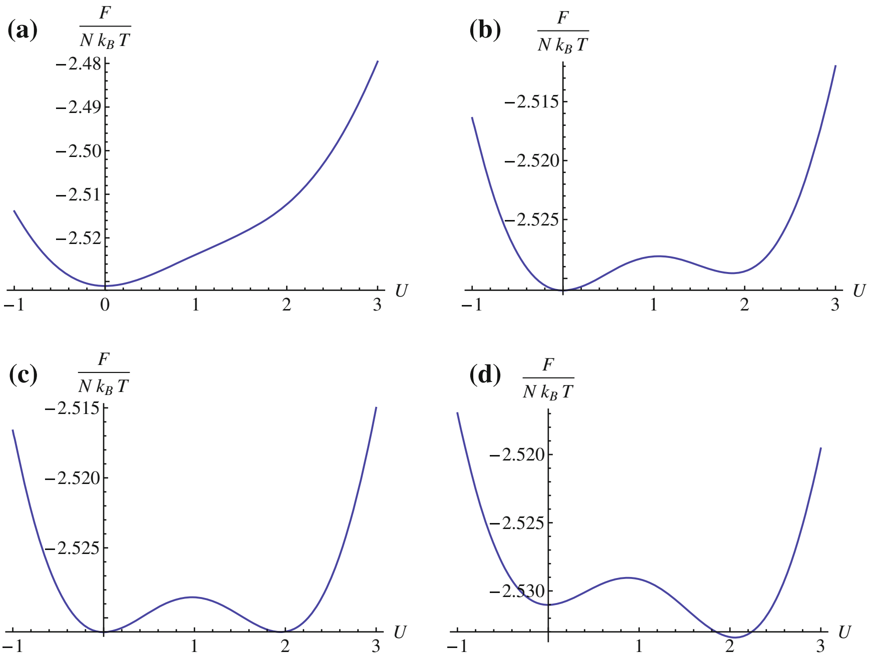


Fig. 10.7 Plots of the Maier-Saupe free energy $F/Nk_B T$ as a function of the effective potential strength U. **a** Far above the isotropic-nematic transition $k_B T/Jq = 0.23$. **b** Slightly above the isotropic-nematic transition $k_B T/Jq = 0.221$. **c** At the the isotropic-nematic transition $k_B T/Jq = 0.2202$. **d** Below the isotropic-nematic transition $k_B T/Jq = 0.219$ (Interactive version at http://www.springer.com/cda/content/document/cda_downloaddocument/Selinger+Interactive+Figures.zip?SGWID=0-0-45-1509169-p177545420.)

As the temperature decreases, in Fig. 10.7b, the free energy has two minima: the stable isotropic minimum and an unstable nematic minimum. At a certain temperature, in Fig. 10.7c, the isotropic and nematic minima become equally deep, and the system has a first-order transition from the isotropic phase to the nematic phase. At a lower temperature, in Fig. 10.7d, the nematic minimum is stable and the isotropic minimum is only metastable.

You will notice that the Maier-Saupe free energy of Fig. 10.7 has the same general form as the Landau-de Gennes free energy of Fig. 10.3. This similarity shows that the macroscopic arguments of Landau-de Gennes theory are actually consistent with microscopic mean-field theory. One advantage of the Maier-Saupe microscopic theory is that it gives specific numerical predictions for the first-order isotropic-nematic transition. It shows that the transition occurs at $k_B T/Jq = 0.2202$, and hence the transition temperature is

$$T_{IN} = 0.2202 \frac{Jq}{k_B}. \tag{10.56}$$

Thus, the theory predicts that the isotropic-nematic transition temperature is proportional to the interaction strength J, and to the number of interacting neighbors q. Just on the nematic side of the transition, the theory predicts that $U = 1.95$. From Eq. (10.51) and Fig. 10.6, this value of U corresponds to a nematic order parameter of $S = 0.429$.

Figure 10.8 shows plots of the Maier-Saupe predictions for the effective potential U and the nematic order parameter S over a broad range of scaled temperature $k_B T/Jq$. Both of these plots show the first-order isotropic nematic transition at $k_B T/Jq = 0.2202$, similar to the Landau-de Gennes prediction of Fig. 10.4. However, there is one important difference: As noted in the previous section, the Landau-de Gennes prediction for S does not reach a maximum at 1, but rather increases without limit as the temperature decreases. This unphysical behavior occurs because the Landau-de Gennes theory is a power series about $S = 0$, and hence is not aware that the maximum possible order parameter is $S = 1$. By contrast, Maier-Saupe theory is *not* a power-series expansion, and it *is* aware that the maximum possible order parameter is $S = 1$. Hence, the Maier-Saupe prediction for S goes to 1 when the temperature goes to 0, as it should.

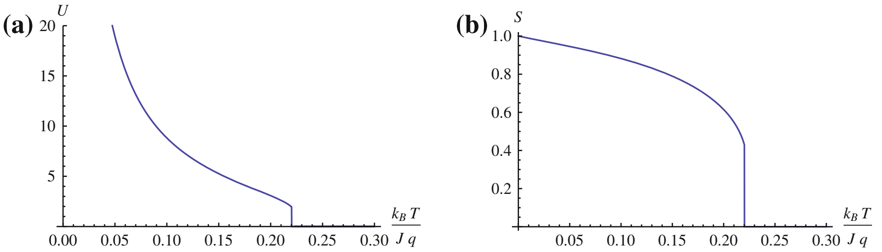


Fig. 10.8 Predictions of Maier-Saupe theory for **a** the effective potential U, and **b** the nematic order parameter S, as functions of scaled temperature $k_B T/Jq$, showing the first-order isotropic-nematic transition

Alternative Mathematical Approach

Let us go back to Eq. 10.46, which gives the Maier-Saupe free energy in terms of the orientational distribution function $\rho(\theta, \phi)$. In the preceding argument, we made a physically motivated assumption (or *ansatz*) for $\rho(\theta, \phi)$ in terms of the single parameter U, and then minimized the free energy over U. You might not believe this assumption for $\rho(\theta, \phi)$. You might ask: Is it really necessary to make any assumption for $\rho(\theta, \phi)$?

The answer is: No, we do not need to make any assumption for $\rho(\theta, \phi)$. As a more general alternative, we can just ask: Of all possible orientational distribution functions, what function $\rho(\theta, \phi)$ gives the minimum free energy? This is a variational calculus problem, as discussed in Chap. 5! We can solve it by calculating the functional derivative of the free energy with respect to $\rho(\theta, \phi)$.

One slight complication in this problem is that we have a constraint: $\rho(\theta, \phi)$ must be normalized, as given by Eq. (10.44). We can implement this constraint using the standard mathematical method of *Lagrange multipliers*. In this method, the minimization equation is

$$\frac{\delta F}{\delta\rho(\theta,\phi)} = \lambda\frac{\delta(\text{constraint})}{\delta\rho(\theta,\phi)} = \lambda\frac{\delta}{\delta\rho(\theta,\phi)}\left[\left(\int_0^{\pi}\sin\theta d\theta\int_0^{2\pi}d\phi\rho(\theta,\phi)\right)-1\right], \tag{10.57}$$

where λ is the Lagrange multiplier. With the free energy of Eq. (10.46), the minimization equation becomes

$$-NqJSP_2(\cos\theta) + Nk_BT\left[1+\log\rho(\theta,\phi)\right] = \lambda. \tag{10.58}$$

The solution is

$$\rho(\theta,\phi) = \exp\left[-1+\frac{\lambda}{Nk_BT}\right]\exp\left[\frac{Jq}{k_BT}SP_2(\cos\theta)\right]. \tag{10.59}$$

We can now use the normalization constraint to determine the Lagrange multiplier λ, and we obtain

$$\rho(\theta,\phi) = \frac{\exp\left[\dfrac{Jq}{k_BT}SP_2(\cos\theta)\right]}{2\pi\displaystyle\int_0^{\pi}\sin\theta d\theta\exp\left[\dfrac{Jq}{k_BT}SP_2(\cos\theta)\right]}. \tag{10.60}$$

Here, S must satisfy the self-consistency equation

$$S = \langle P_2(\cos\theta)\rangle = \frac{\displaystyle\int_0^{\pi}\sin\theta d\theta P_2(\cos\theta)\exp\left[\dfrac{Jq}{k_BT}SP_2(\cos\theta)\right]}{\displaystyle\int_0^{\pi}\sin\theta d\theta\exp\left[\dfrac{Jq}{k_BT}SP_2(\cos\theta)\right]}. \tag{10.61}$$

You should notice that the result (10.60) has exactly the same form as our previous assumption (10.49) for the orientational distribution function! Furthermore, the self-consistency equation (10.61) is equivalent to our previous minimization condition (10.55). Hence, our physically motivated approach for minimizing the free energy really is justified by variational calculus.

10.5 Onsager Theory

At this point, let us briefly consider another microscopic model for the isotropic-nematic transition, Onsager theory, which makes different assumptions about the fundamental interactions between particles.

Onsager theory assumes that the liquid crystal is composed of *hard rods*. These rods have no interaction other than excluded-volume repulsion, which prevents them from overlapping. The interaction potential between rods i and j is

$$V_{ij} = \begin{cases} 0 & \text{if no overlap,} \\ \infty & \text{if overlap.} \end{cases} \tag{10.62}$$

Hence, these particles are anisotropic versions of the hard spheres discussed in Sect. 8.3. This is generally *not* a good approximation for molecules that form liquid-crystal phases, like 5CB. However, it is a good approximation for colloidal particles in solution. One example of rod-like colloidal particles is viruses; other examples could be synthetic particles fabricated by experimenters. For such particles, the question is: When you put many hard rods in solution, what phase will they form—isotropic or nematic?

To address this question, we first need to determine how much volume is excluded by the interaction between rods. Suppose each rod has length L and diameter D, with $L \gg D$. Let us consider two rods with a relative orientation θ, as shown in Fig. 10.9. Here, the first rod excludes a certain amount of volume to the second rod—there are certain places where the second rod *cannot* go, because it would overlap with the first rod. The center-of-mass of the second rod can go anywhere in the system *except* the forbidden zone, which is shown by the pink box in the figure. The forbidden zone is a parallelepiped with height L, width $L|\sin\theta|$, and depth $2D$ (plus some small corrections of order D/L, which are unimportant in the limit of long narrow rods). As a result, the excluded volume is

$$V_{\text{excluded}} = 2DL^2|\sin\theta|. \tag{10.63}$$

The important point in this equation is that the excluded volume depends on the relative orientation between the rods. It is smallest when the rods are parallel (or antiparallel) and largest when the rods are perpendicular.

The next step is to determine the free energy of a system with many rods. Because there is no interaction energy other than excluded volume, the expectation value

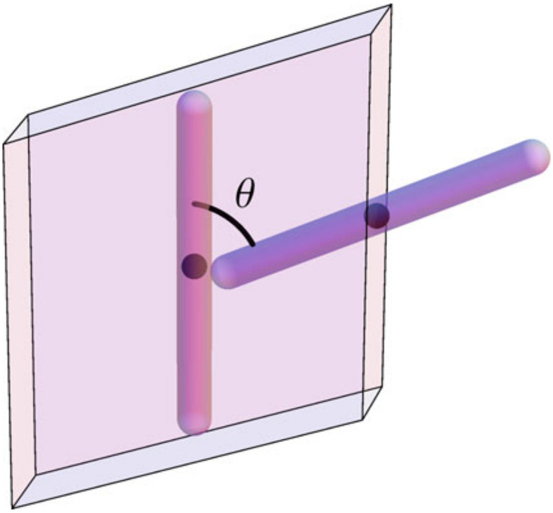


Fig. 10.9 Two hard rods interacting through the excluded-volume interaction, in the Onsager theory of the isotropic-nematic transition. The *pink* parallelepiped shows the volume that is forbidden to the second rod because of the first rod, at fixed relative orientation θ (Interactive version at http://www.springer.com/cda/content/document/cda_downloaddocument/Selinger+Interactive+Figures.zip?SGWID=0-0-45-1509169-p177545420.)

of the energy is $\langle E \rangle = 0$, and the free energy is just $F = -TS_{\text{entropy}}$. As you recall from van der Waals theory in Chap. 3, excluded volume *reduces* the number of allowed configurations of the system, and hence *reduces* the entropy and *increases* the free energy. As a result, the excluded-volume contribution to the free energy favors alignment of the particles in a nematic phase.

This favored alignment might seem surprising, because we normally expect that entropy will favor disorder rather than order. The best way to understand this result is to think about orientational entropy versus positional entropy. Orientational entropy favors an isotropic phase, with the greatest possible disorder in the rod orientations. By contrast, positional entropy favors a nematic phase, so that the rods will be aligned, will have the minimum excluded volume, and will have the greatest possible number of available positions. This competition between orientational and positional entropy is similar to the competition between long-range and short-range entropy in hard-sphere crystallization in Sect. 8.3.

Onsager did a calculation to compare the free energies of the isotropic and nematic phases, with the excluded-volume interaction of Eq. (10.63). In this calculation, the phase transition does not depend on temperature because the interaction is purely entropic; there is no competition between energy and entropy. Rather, the transition depends on the concentration c of rods per unit volume, compared with the parameter DL^2 in the interaction. The isotropic phase occurs for concentration $c < 4.25/(DL^2)$, and the nematic phase occurs for $c > 5.71/(DL^2)$. The calculation also shows that the order parameter just on the nematic side of the transition is $S = 0.84$, which is substantially larger than the order parameter predicted by Maier-

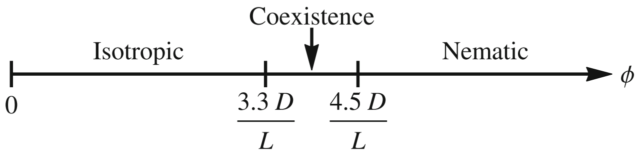


Fig. 10.10 Phase diagram for hard rods as a function of one variable, the volume fraction ϕ

Saupe theory. In other words, there is a bigger first-order discontinuity in Onsager theory than in Maier-Saupe theory.

You should notice that the phase transition depends on the product DL^2, unlike the volume of a rod, which depends on D^2L. Based on this contrast, it is interesting to re-express the result in terms of the volume fraction

$$\phi = \frac{\text{volume in rods}}{\text{total volume}} = \left(\frac{\text{number of rods}}{\text{total volume}}\right)(\text{volume of 1 rod}) = (c)\left(\frac{\pi D^2 L}{4}\right). \tag{10.64}$$

In terms of volume fraction, the isotropic phase occurs for $\phi < 3.3D/L$, and the nematic phase occurs for $\phi > 4.5D/L$. Hence, the volume fraction required for a nematic phase is inversely proportional to the aspect ratio L/D of the rods. If the rods are long and thin, it is easy to form a nematic phase, which will occur at a low volume fraction. If the rods are short and fat, it is difficult to form a nematic phase.

Figure 10.10 shows the phase diagram in terms of the single variable ϕ, with the isotropic phase for $\phi < 3.3D/L$ and the nematic phase for $\phi > 4.5D/L$. If the system is prepared with a volume fraction of rods between these two values, then it will exhibit two-phase coexistence between isotropic and nematic phases, analogous to the liquid-gas coexistence at fixed volume in Chap. 3. The fraction of each phase will be given by the level rule.

As a matter of terminology, if the isotropic-nematic transition is controlled by concentration, the system is called a *lyotropic* liquid crystal. By contrast, if the isotropic-nematic transition is controlled by temperature, the system is called a *thermotropic* liquid crystal. Hence, Onsager theory for hard rods in solution is an example of a lyotropic liquid crystal, and Maier-Saupe theory for interacting molecules is an example of a thermotropic liquid crystal.

10.6 Elasticity of Nematic Order

In Sects. 10.3–10.5, we saw that molecular interactions determine the magnitude S of nematic order, but they do not determine the orientation $\hat{\boldsymbol{n}}$ of nematic order. Indeed, molecular interactions do not care about the direction of nematic order. The free energy depends on S, but it does not depend on the orientation of the director in 3D space. The liquid crystal can form a nematic phase with $\hat{\boldsymbol{n}}$ pointing in any orientation, and the system randomly chooses an orientation. This is the essential concept of spontaneous symmetry breaking!

Although molecular interactions do not care about the direction of nematic order, they do care about spatial variations in the direction. If $\hat{\boldsymbol{n}}$ is gradually changing as a function of position, then the molecules are not as well aligned as if $\hat{\boldsymbol{n}}$ were uniform. As a result, the free energy will be higher. This extra free energy associated with variations in $\hat{\boldsymbol{n}}$ is called the *elastic free energy,* or the *Frank free energy.*

To understand the elastic free energy, we must use the concept of fields, as in Chap. 6. For liquid crystals, we can consider $Q_{\alpha\beta}(\boldsymbol{r})$ as a *tensor field,* which gives the nematic order for molecules near the position $\boldsymbol{r}$. How near? Well, that is the same argument as in Chap. 6. We must do a coarse-grained average of a local group of molecules, with a length scale much bigger than the size of a molecule and much smaller than the scale on which physical properties vary. This definition of a field is possible because molecules are very small compared with typical experiments.

Once we have $Q_{\alpha\beta}(\boldsymbol{r})$ as a field, we can define the scalar order parameter field $S(\boldsymbol{r})$ and the director field $\hat{\boldsymbol{n}}(\boldsymbol{r})$, so that

$$Q_{\alpha\beta}(\boldsymbol{r}) = S(\boldsymbol{r})\left[\frac{3}{2}n_\alpha(\boldsymbol{r})n_\beta(\boldsymbol{r}) - \frac{1}{2}\delta_{\alpha\beta}\right]. \tag{10.65}$$

In most systems, $S(\boldsymbol{r})$ is strongly determined by the molecular interactions at a particular temperature (in a thermotropic liquid crystal) or concentration (in a lyotropic liquid crystal). It would cost a huge free energy to change $S(\boldsymbol{r})$ to anything other than the value determined by Sects. 10.3–10.5. Hence, $S(\boldsymbol{r})$ usually does not depend on position very much. For an excellent approximation, we can regard it as a constant independent of position, and write

$$Q_{\alpha\beta}(\boldsymbol{r}) = S\left[\frac{3}{2}n_\alpha(\boldsymbol{r})n_\beta(\boldsymbol{r}) - \frac{1}{2}\delta_{\alpha\beta}\right]. \tag{10.66}$$

By contrast, $\hat{\boldsymbol{n}}(\boldsymbol{r})$ is not determined by molecular interactions—it can easily vary as a function of position, and hence we must keep it as a field.

Next we must ask: How do spatial variations of $Q_{\alpha\beta}(\boldsymbol{r})$ or $\hat{\boldsymbol{n}}(\boldsymbol{r})$ affect the free energy? To answer this question, let us return to the Landau-de Gennes free energy density of Eq. (10.20). When $Q_{\alpha\beta}(\boldsymbol{r})$ is a function of position, we must add further terms to the free energy involving gradients of $Q_{\alpha\beta}(\boldsymbol{r})$. These terms must satisfy all the appropriate symmetries, so that the free energy is still a scalar. The simplest such term is $\frac{1}{2}L(\partial_\gamma Q_{\alpha\beta})(\partial_\gamma Q_{\alpha\beta})$, where L is an arbitrary elastic coefficient. This term shows that variations of any component $Q_{\alpha\beta}$ in any direction r_γ cost free energy. It is not the only elastic term permitted by symmetry, but let us also consider just that term for now. The Landau-de Gennes free energy density then becomes

$$\begin{aligned} f =& f_0 + \frac{1}{2}AQ_{\alpha\beta}Q_{\alpha\beta} + \frac{1}{3}BQ_{\alpha\beta}Q_{\beta\gamma}Q_{\gamma\alpha} + \frac{1}{3}C_1(Q_{\alpha\beta}Q_{\alpha\beta})^2 \\ &+ \frac{1}{3}C_2Q_{\alpha\beta}Q_{\beta\gamma}Q_{\gamma\delta}Q_{\delta\alpha} + \frac{1}{2}L(\partial_\gamma Q_{\alpha\beta})(\partial_\gamma Q_{\alpha\beta}), \end{aligned} \tag{10.67}$$

and the full free energy is the integral of the free energy density over the whole system

$$F = \int d^3r \Big[f_0 + \frac{1}{2} A Q_{\alpha\beta} Q_{\alpha\beta} + \frac{1}{3} B Q_{\alpha\beta} Q_{\beta\gamma} Q_{\gamma\alpha} + \frac{1}{3} C_1 (Q_{\alpha\beta} Q_{\alpha\beta})^2 + \frac{1}{3} C_2 Q_{\alpha\beta} Q_{\beta\gamma} Q_{\gamma\delta} Q_{\delta\alpha} + \frac{1}{2} L (\partial_\gamma Q_{\alpha\beta})(\partial_\gamma Q_{\alpha\beta}) \Big]. \tag{10.68}$$

Now let us express the elastic free energy in terms of the director field $\hat{\boldsymbol{n}}(\boldsymbol{r})$. By substituting the expression (10.66) into the free energy (10.68), we obtain

$$F = \int d^3r \left[f_{\text{uniform}} + \frac{9}{8} L S^2 \left(n_\alpha \partial_\gamma n_\beta + n_\beta \partial_\gamma n_\alpha \right) \left(n_\alpha \partial_\gamma n_\beta + n_\beta \partial_\gamma n_\alpha \right) \right]. \tag{10.69}$$

Here, the first term f_{uniform} is the Landau-de Gennes free energy density of the uniform liquid crystal. The important point is that it does not depend on $\hat{\boldsymbol{n}}(\boldsymbol{r})$, and hence it is a constant, which we can neglect. The second term can be simplified when we use the constraint that $\hat{\boldsymbol{n}}(\boldsymbol{r})$ is a unit vector:

$$n_\alpha n_\alpha = 1, \tag{10.70}$$

and hence

$$n_\alpha \partial_\gamma n_\alpha = \frac{1}{2} \partial_\gamma (n_\alpha n_\alpha) = \frac{1}{2} \partial_\gamma (1) = 0. \tag{10.71}$$

Thus, the elastic free energy becomes

$$F = \int d^3r \left[\frac{1}{2} K \left(\partial_\gamma n_\alpha \right) \left(\partial_\gamma n_\alpha \right) \right], \tag{10.72}$$

where $K = \frac{9}{2} L S^2$.

Equation (10.72) is the simplest version of the Frank free energy. It shows that variations of any director component n_α in any direction r_γ cost free energy. The coefficient K is the simplest version of a Frank elastic constant. Here, we have found that K is proportional to the tensor elastic constant L times S^2. Hence, as the temperature decreases and the nematic phase becomes more ordered, we would expect K to increase proportional to S^2.

Although Eq. (10.72) is a reasonable first approximation, it is not good enough for modeling liquid crystals in detail, nor for designing liquid-crystal devices. The problem with this expression is that it implies that all variations in the director field cost the same amount of free energy, regardless of the direction of the variation with respect to the director itself. Nematic liquid crystals are more anisotropic than that: There are different free energy costs for variations in different directions, and we need different Frank constants to model them.

To develop an improved version of the elastic free energy for nematic liquid crystals, there are two possible approaches. First, we can go back to the Landau-de Gennes expression for the free energy density in terms of $Q_{\alpha\beta}$ and construct more derivative terms. For example, we can make terms like $(\partial_\alpha Q_{\alpha\gamma})(\partial_\beta Q_{\beta\gamma})$, which contracts the indices in different ways, and $Q_{\gamma\delta}(\partial_\gamma Q_{\alpha\beta})(\partial_\delta Q_{\alpha\beta})$, which goes to higher order in the tensor order parameter. We can then express all these terms using the director field. Alternatively, we can stop thinking about the order tensor at all, and just work directly with the director field. From $\hat{\boldsymbol{n}}(\boldsymbol{r})$ and its derivatives, we can identify the modes of distortion that have different free energy costs. This second approach has two advantages: It is generally easier, and it does not require any assumption that the scalar order parameter S is small. For those reasons, we will follow the second approach here.

To classify the modes of distortion in a nematic liquid crystal, we must ask two questions: What is the orientation of the local average $\hat{\boldsymbol{n}}(\boldsymbol{r})$ with respect to the gradient direction—parallel or perpendicular to the gradient direction? Which component of $\hat{\boldsymbol{n}}(\boldsymbol{r})$ is varying—parallel or perpendicular to the gradient direction? Based on the answers to those questions, we can identify the three distortion modes shown in Fig. 10.11:

1. *Splay:* If the local average $\hat{\boldsymbol{n}}(\boldsymbol{r})$ is *perpendicular* to the gradient direction, and the varying component of $\hat{\boldsymbol{n}}(\boldsymbol{r})$ is *parallel* to the gradient direction, then the distortion is called splay. Splay is characterized by the vector $\hat{\boldsymbol{n}}(\nabla \cdot \hat{\boldsymbol{n}})$. This splay vector is

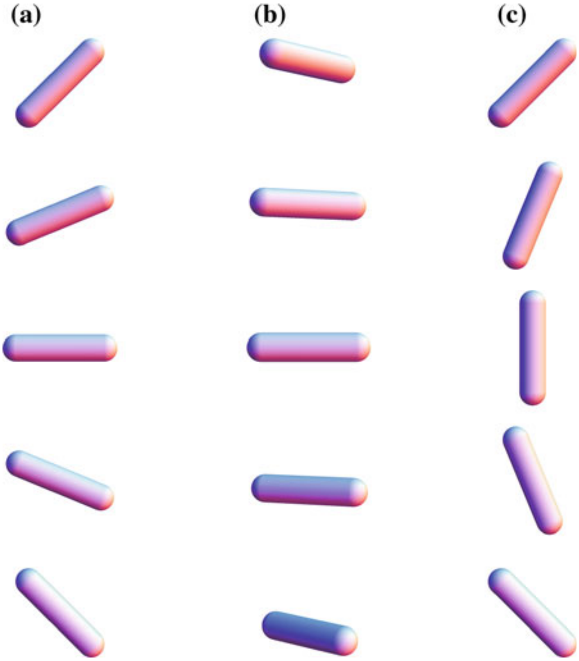


Fig. 10.11 Visualization of the three nematic modes of distortion: **a** Splay. **b** Twist. **c** Bend (Interactive version at http://www.springer.com/cda/content/document/cda_downloaddocument/Selinger+Interactive+Figures.zip?SGWID=0-0-45-1509169-p177545420.)

first-order in the gradient operator, so it shows gradual variations in the director orientation. It is even in $\hat{\boldsymbol{n}}$, as required by symmetry because $\hat{\boldsymbol{n}}$ and $-\hat{\boldsymbol{n}}$ represent the same physical state. The free energy density associated with splay is given by $f_{\text{splay}} = \frac{1}{2}K_1|\hat{\boldsymbol{n}}(\nabla \cdot \hat{\boldsymbol{n}})|^2 = \frac{1}{2}K_1(\nabla \cdot \hat{\boldsymbol{n}})^2$, which is a scalar. Here, K_1 is the Frank elastic constant for splay.

2. *Twist:* If the local average $\hat{\boldsymbol{n}}(\boldsymbol{r})$ is *perpendicular* to the gradient direction, and the varying component of $\hat{\boldsymbol{n}}(\boldsymbol{r})$ is also *perpendicular* to the gradient direction, then the distortion is called twist. Twist is characterized by $\hat{\boldsymbol{n}} \cdot (\nabla \times \hat{\boldsymbol{n}})$, which is a pseudoscalar. It is reasonable that twist should be a pseudoscalar because this distortion has handedness; it changes sign under inversion. Like the splay vector, the twist pseudoscalar is first-order in the gradient operator and is even in $\hat{\boldsymbol{n}}$. The free energy density associated with twist is given by $f_{\text{twist}} = \frac{1}{2}K_2[\hat{\boldsymbol{n}} \cdot (\nabla \times \hat{\boldsymbol{n}})]^2$, which is a proper scalar; it does not change sign under inversion. Here, K_2 is the Frank elastic constant for twist.
3. *Bend:* If the local average $\hat{\boldsymbol{n}}(\boldsymbol{r})$ is *parallel* to the gradient direction, then the varying component of $\hat{\boldsymbol{n}}(\boldsymbol{r})$ must be *perpendicular* to the gradient direction because $\hat{\boldsymbol{n}}(\boldsymbol{r})$ is a unit vector. In this case, the distortion is called bend, and it is characterized by the vector $\hat{\boldsymbol{n}} \times (\nabla \times \hat{\boldsymbol{n}})$. Again, the bend vector is first-order in the gradient operator and is even in $\hat{\boldsymbol{n}}$. The free energy density associated with bend is given by the scalar $f_{\text{bend}} = \frac{1}{2}K_3|\hat{\boldsymbol{n}} \times (\nabla \times \hat{\boldsymbol{n}})|^2$, where K_3 is the Frank elastic constant for bend.

Putting these three modes together, the Frank free energy for elastic distortions in a nematic liquid crystal becomes

$$F = \int d^3r \left[\frac{1}{2}K_1(\nabla \cdot \hat{\boldsymbol{n}})^2 + \frac{1}{2}K_2[\hat{\boldsymbol{n}} \cdot (\nabla \times \hat{\boldsymbol{n}})]^2 + \frac{1}{2}K_3|\hat{\boldsymbol{n}} \times (\nabla \times \hat{\boldsymbol{n}})|^2 \right]. \tag{10.73}$$

This is the form that is normally used in liquid-crystal research and development.

From Eq. (10.73), we can see that the Frank constants all have dimensions of energy/length, so they are generally reported in Newtons. In most nematic liquid crystals, they are all around 10^{-11} N. This order of magnitude can be understood through the following argument: Based on dimensional analysis, we would expect the Frank constants to be roughly a typical energy divided by a typical length. For a nematic liquid crystal, the typical energy is given by the intermolecular interaction energy, which is (by Maier-Saupe theory) comparable to $k_B T_{IN}$, where T_{IN} is the isotropic-nematic transition temperature. We have $k_B = 1.38 \times 10^{-23}$ J/K and might estimate $T_{IN} \approx 300$ K, which gives an interaction energy around 4×10^{-21} J. The typical length is the distance between molecules, which is around 10^{-9} m. From the ratio, we estimate that the Frank constants should be around 4×10^{-12} N, which is roughly consistent with typical experimental numbers.

In general, the three Frank constants are not equal to each other, but they are not very different. You might ask: What if they were equal to each other, with $K_1 = K_2 = K_3 \equiv K$? Would the general Frank free energy of Eq. (10.73) reduce to our first approximation of Eq. (10.72)?

The answer is: Almost, but not exactly! Our expressions are actually related by

$$\frac{1}{2}K\left(\partial_\gamma n_\alpha\right)\left(\partial_\gamma n_\alpha\right) = \frac{1}{2}K_1(\nabla\cdot\hat{\boldsymbol{n}})^2 + \frac{1}{2}K_2[\hat{\boldsymbol{n}}\cdot(\nabla\times\hat{\boldsymbol{n}})]^2 + \frac{1}{2}K_3|\hat{\boldsymbol{n}}\times(\nabla\times\hat{\boldsymbol{n}})|^2 + \frac{1}{2}K_{24}\nabla\cdot[(\hat{\boldsymbol{n}}\cdot\nabla)\hat{\boldsymbol{n}} - \hat{\boldsymbol{n}}(\nabla\cdot\hat{\boldsymbol{n}})], \tag{10.74}$$

in the limit where $K_1 = K_2 = K_3 = K_{24} \equiv K$. The final term, proportional to K_{24}, is one extra mode of distortion called *saddle-splay*. In the saddle-splay mode, the planes perpendicular to $\hat{\boldsymbol{n}}(\boldsymbol{r})$ are shaped like saddles, with negative Gaussian curvature, leading to a director field that is not easy to draw in a single figure. From Eq. (10.74), we can see that the saddle-splay term is the divergence of a vector field. By the divergence theorem of vector calculus, the volume integral of the saddle-splay term can be reduced to a surface integral over the boundary. For this reason, the saddle-splay term is often called a surface elastic term. In many typical liquid-crystal systems, the director field on the surface is fixed by some anchoring conditions. In these cases, the saddle-splay term is a constant, independent of the director field in the interior of the cell, so we can neglect it. Hence, it is not commonly used in studies of liquid-crystal devices. However, there are certain unusual situations in which the director field on the boundaries can vary, or in which the liquid crystal can change its boundaries; in these situations, the saddle-splay term must be considered.

10.7 Frederiks Transition

In the previous section, we saw that molecular interactions do not care about the direction of nematic order, but only about spatial variations in the direction. However, it is possible to apply *symmetry-breaking fields,* which *do* care about the direction of nematic order. Two examples of symmetry-breaking fields are *electric* and *magnetic* fields, which tend to favor alignment of nematic order either parallel or perpendicular to the fields (depending on the particular liquid-crystal material). Another example is *surface anchoring;* at the surface of a liquid-crystal cell, interactions of the molecules with the surface can favor alignment in some direction with respect to the surface.

Now we can consider one of the most common problems in liquid-crystal science and technology: Suppose we have a system with different symmetry-breaking fields that compete with each other. How does the liquid crystal respond to these competing influences? In other words, what director field $\hat{\boldsymbol{n}}(\boldsymbol{r})$ minimizes the free energy?

Let us investigate a classic example of this problem, the *Frederiks transition.* (It is sometimes written as the Freedericks or Freédericksz transition, based on an archaic method for writing Russian names in German.) In this problem, we suppose that there are two competing symmetry-breaking fields—surface anchoring and a magnetic (or electric) field—and we look for the director field that arises from this competition.

First, suppose we have a semi-infinite liquid crystal, as shown in Fig. 10.12. On the surface at $x = 0$, there is very strong anchoring, which requires the director to align in the z-direction. Everywhere in the bulk of the liquid crystal, for $x > 0$, there is a magnetic field, which favors alignment of the director in the y-direction. To minimize the free energy, the director field must adopt a configuration where it is aligned with the z-direction near the surface, and then gradually twists into the y-direction into the bulk. The question is then: How far does the surface alignment extend into the bulk?

To model this problem, we represent the director field as

$$\hat{\boldsymbol{n}}(x) = (0, \sin\theta(x), \cos\theta(x)). \tag{10.75}$$

The surface anchoring provides the boundary condition that $\theta(0) = 0$. There are now two contributions to the free energy. The first contribution is the Frank free energy of Eq. (10.73). When we insert our form (10.75), it simplifies to

$$F_{\text{Frank}} = A \int dx \left[\frac{1}{2} K_2 \left(\frac{d\theta}{dx} \right)^2 \right], \tag{10.76}$$

where A is the cross-sectional area in the yz-plane; note that this distortion involves only twist. The second contribution is the interaction of the magnetic field of the liquid crystal, which can be written as

$$F_{\text{magnetic}} = A \int dx \left[-\frac{1}{2} \frac{\Delta\chi}{\mu_0} (\boldsymbol{H} \cdot \hat{\boldsymbol{n}}(x))^2 \right] = A \int dx \left[-\frac{1}{2} \frac{\Delta\chi}{\mu_0} H^2 \sin^2\theta(x) \right], \tag{10.77}$$

where $\boldsymbol{H}$ is the magnetic field and $\Delta\chi$ is the diamagnetic anisotropy of the liquid crystal. (It is related to the maximum diamagnetic anisotropy discussed in Sect. 10.3 by $\Delta\chi = \Delta\chi_{\max} S$.) Let us assume that we have a liquid crystal with $\Delta\chi > 0$, so that the field favors alignment in the y-direction, with $\theta = \pm\pi/2$. The total free energy is then

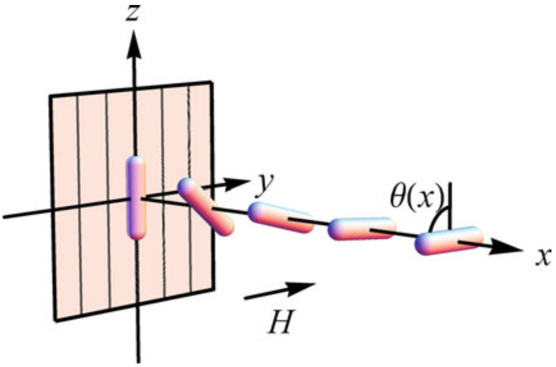


Fig. 10.12 Semi-infinite liquid crystal, with strong surface anchoring that requires $\theta = 0$ at $x = 0$, and a magnetic field that favors $\theta = \pm\pi/2$ in the bulk (Interactive version at http://www.springer.com/cda/content/document/cda_downloaddocument/Selinger+Interactive+Figures.zip?SGWID=0-0-45-1509169-p177545420.)

$$F = \frac{1}{2} A \int dx \left[K_2 \left(\frac{d\theta}{dx} \right)^2 - \frac{\Delta\chi}{\mu_0} H^2 \sin^2 \theta(x) \right]. \tag{10.78}$$

We want to determine what function $\theta(x)$ minimizes the free energy. Of course, this is a variational calculus problem. To solve it, we calculate the functional derivative of the free energy with respect to $\theta(x)$, and set this functional derivative equal to zero, to obtain the Euler-Lagrange equation

$$\frac{\delta F}{\delta\theta(x)} = -A \left[K_2 \frac{d^2\theta}{dx^2} + \frac{\Delta\chi}{\mu_0} H^2 \sin\theta(x) \cos\theta(x) \right] = 0. \tag{10.79}$$

This equation is often written as

$$\frac{d^2\theta}{dx^2} = -\frac{1}{\xi^2} \sin\theta(x) \cos\theta(x), \tag{10.80}$$

where

$$\xi = \frac{1}{H} \sqrt{\frac{K_2 \mu_0}{\Delta\chi}} \tag{10.81}$$

is called the *magnetic coherence length.* Through some mathematical manipulation, we can show that the solution of this differential equation (with the boundary conditions that $\theta = 0$ at $x = 0$, and $\theta =$ constant at $x \to \infty$) is

$$\theta(x) = \pm \sin^{-1} \left(\tanh \frac{x}{\xi} \right). \tag{10.82}$$

This solution is plotted in Fig. 10.13. From this plot, we can see that the magnetic coherence length ξ gives the length scale over which the surface alignment extends into the bulk. If the field is low, then ξ is large, and the surface alignment extends far into the bulk. If the field is high, then ξ is small, so the alignment is mainly dominated

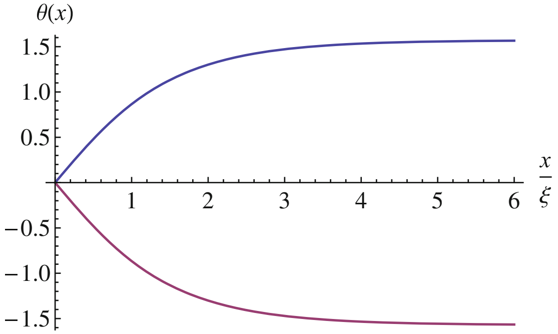


Fig. 10.13 Solution for liquid-crystal alignment in a semi-infinite geometry, with strong surface anchoring that requires $\theta = 0$ at $x = 0$, and a magnetic field that favors $\theta = \pm\pi/2$ in the bulk

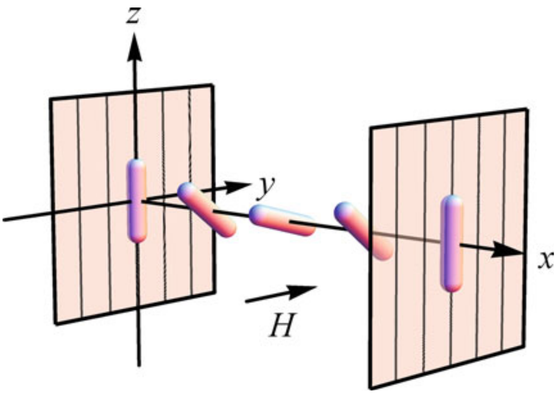


Fig. 10.14 Finite liquid-crystal cell, with strong surface anchoring that requires $\theta = 0$ at $x = 0$ and $x = d$, and a magnetic field that favors $\theta = \pm\pi/2$ in the interior (Interactive version at http://www.springer.com/cda/content/document/cda_downloaddocument/Selinger+Interactive+Figures.zip?SGWID=0-0-45-1509169-p177545420.)

by the field, with only a small surface region. The ± sign indicates that the system can randomly twist to the right or to the left.

For an alternative version of this problem, suppose we have a *finite* liquid-crystal cell, as shown in Fig. 10.14, with aligning surfaces at *both* $x = 0$ and $x = d$ that anchor the director along the z-direction. In the interior, the magnetic field again favors alignment along the y-direction. We would now like to calculate the profile $\theta(x)$ for $0 < x < d$. This finite problem has the same Euler-Lagrange equation as the semi-infinite problem, but the boundary conditions are different: We must now find a solution of the differential equation that satisfies $\theta(0) = \theta(d) = 0$. In general, this problem cannot be solved exactly, but it can certainly be solved numerically.

Figure 10.15 shows a series of numerical solutions for different values of d/ξ, the ratio of the cell thickness to the magnetic coherence length. When d/ξ is small (i.e., when the magnetic field is small), the solution is exactly $\theta(x) = 0$. In this case, the elastic free energy cost of director twist is so great, compared with the small gain in magnetic free energy, that the system does not twist at all. When d/ξ becomes large enough (i.e., when the magnetic field becomes large enough), past a certain threshold, the system begins to twist. In this case, we can see that $\theta(x)$ becomes slightly nonzero in the interior, so that the director can partially align with the magnetic field. When

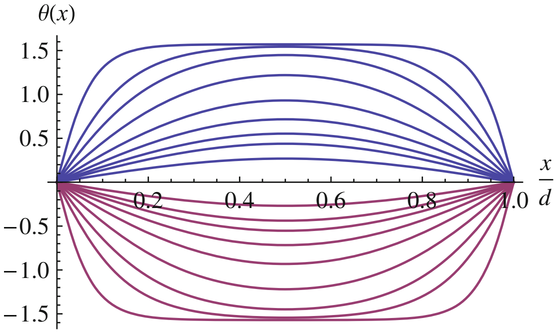


Fig. 10.15 Numerical solution for liquid-crystal alignment in a finite cell, for several values of d/ξ, the ratio of the cell thickness to the magnetic coherence length. From smallest to largest absolute value, the solutions are for $d/\xi = 3.2$, 3.3, 3.4, 3.6, 4, 5, 7, 10, and 20

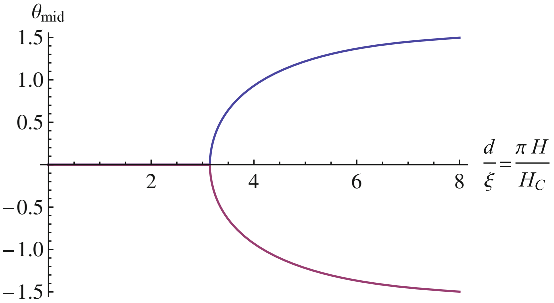


Fig. 10.16 Numerical solution for $\theta_{\rm mid} = \theta(d/2)$, the orientation at the middle of the cell, as function of the ratio d/ξ or equivalently $\pi H/H_C$

d/ξ becomes very large (i.e. when the magnetic field becomes very large), the director almost goes to $\theta(x) = \pm\pi/2$ in the interior, as favored by the magnetic field. It only deviates from the field direction in small surface regions of size ξ. There are both positive and negative solutions, because the system can randomly twist to the right or to the left in the interior.

One way to characterize the solution is by the angle $\theta_{\rm mid} = \theta(d/2)$ at the middle of the cell. Figure 10.16 shows the numerical solution for $\theta_{\rm mid}$ as a function of d/ξ. This plot looks just like the plot of the Ising order parameter as a function of temperature in Fig. 2.6a! At a critical threshold of d/ξ, the liquid crystal has a second-order transition, called the Frederiks transition. Below the critical threshold, we have $\theta(x) = 0$ throughout the cell, and hence $\theta_{\rm mid} = 0$. Above the critical threshold, the director field begins to twist, and $\theta_{\rm mid}$ characterizes the magnitude of the twist distortion. Hence, $\theta_{\rm mid}$ acts as an order parameter for the Frederiks transition.

Like the Ising transition, the Frederiks transition is a symmetry-breaking transition, because $\theta_{\rm mid}$ randomly becomes positive or negative, meaning that the director field randomly twists to the right or to the left. Unlike the Ising transition, the Frederiks transition does *not* involve a competitition between energy and entropy. Rather, it is driven by the competition between elastic free energy and magnetic free energy.

To understand the behavior near the threshold, consider the functional form for $\theta(x)$ shown in Fig. 10.15. Near the threshold, $\theta(x)$ can be well approximated by

$$\theta(x) = \theta_{\rm mid} \sin\frac{\pi x}{d}. \tag{10.83}$$

The deviation from surface alignment is small, so that $\theta_{\rm mid} \ll 1$. We put that approximate form into the free energy of Eq. (10.78), and integrate over $x = 0$ to d, to obtain

$$F = \frac{\pi^2 K_2}{4d}\theta_{\rm mid}^2 - \frac{d\Delta\chi H^2}{4\mu_0}[1 - J_0(2\theta_{\rm mid})], \tag{10.84}$$

where J_0 is a Bessel function. We then expand the free energy as a power series in $\theta_{\rm mid}$ to obtain

$$\begin{aligned} F &= \left(\frac{\pi^2 K_2}{4d} - \frac{d\Delta\chi H^2}{4\mu_0} \right) \theta_{\text{mid}}^2 + \frac{d\Delta\chi H^2}{16\mu_0} \theta_{\text{mid}}^4 + \cdots \\ &= \frac{K_2}{4d} \left(\pi^2 - \frac{d^2}{\xi^2} \right) \theta_{\text{mid}}^2 + \frac{dK_2}{16\xi^2} \theta_{\text{mid}}^4 + \cdots . \end{aligned} \tag{10.85}$$

This mathematical form for the free energy is just like the free energy of the Ising model. By analogy with the Ising model, the Frederiks transition occurs when the coefficient of the quadratic term passes through zero, which occurs at

$$\frac{d}{\xi} = \pi, \tag{10.86}$$

or equivalently at the critical field

$$H_C = \frac{\pi}{d} \sqrt{\frac{K_2 \mu_0}{\Delta\chi}}. \tag{10.87}$$

For $H < H_C$, we find $\theta_{\text{mid}} = 0$. For H slightly greater than H_C, θ_{mid} increases as

$$\theta_{\text{mid}} = \pm\sqrt{2\left(1 - \frac{\pi^2\xi^2}{d^2}\right)} \propto \pm(H - H_C)^{1/2}. \tag{10.88}$$

This relation is equivalent to the scaling of the Ising order parameter below the critical temperature with the exponent $\beta = \frac{1}{2}$.

For one more variation on the Frederiks problem, we can intentionally break the symmetry between the two possible directions of twist. To do this, we rotate the surface alignment direction with respect to the magnetic field direction, so that they are not exactly perpendicular, as shown in Fig. 10.17. Suppose that the magnetic field is still in the y-direction, but the surface alignment is $(0, \sin\theta_{\text{pre}}, \cos\theta_{\text{pre}})$. Here, θ_{pre}

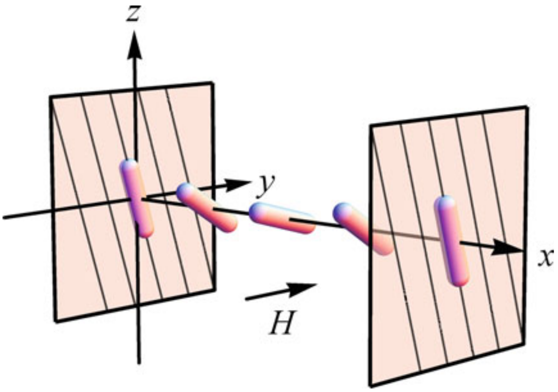


Fig. 10.17 Finite liquid-crystal cell with strong surface anchoring at a pre-tilt angle θ_{pre}, which is *not* perpendicular to the magnetic field direction (Interactive version at http://www.springer.com/cda/content/document/cda_downloaddocument/Selinger+Interactive+Figures.zip?SGWID=0-0-45-1509169-p177545420.)

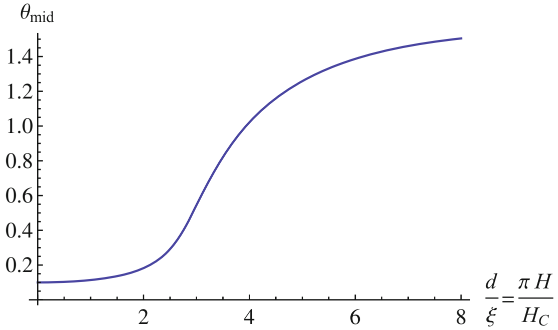


Fig. 10.18 Numerical solution for the orientation θ_{mid} in the middle of the cell, for $\theta_{\text{pre}} = 0.1$ radians

is a *pre-tilt* angle, which biases the director field toward twist in one direction, rather than the opposite direction.

Figure 10.18 shows a numerical solution for the orientation θ_{mid} in the middle of the cell, for the example $\theta_{\text{pre}} = 0.1$ radians. This plot is quite different from the previous case with no pre-tilt in Fig. 10.16. With a pre-tilt, there is no critical threshold! Rather, we see a smooth evolution of θ_{mid} as a function of magnetic field (or d/ξ). For small field, θ_{mid} is slightly shifted from θ_{pre}. As the field increases, θ_{mid} gradually increases and eventually approaches $\pi/2$. Furthermore, there is only one solution θ_{mid} that minimizes the free energy; we do not have two minima. This behavior occurs because we do not need a transition to spontaneously break the symmetry between positive and negative θ_{mid}; the symmetry is already broken by the pre-tilt! In this respect, the pre-tilt is analogous to the symmetry-breaking field h on the Ising model, which induces order even if $T > T_C$, as shown in Fig. 2.6b.

In the literature, there are many other variations on the Frederiks problem. The applied field can be either magnetic or electric. Furthermore, the relative orientations of surface anchoring and field alignment can be chosen so that the director distortion is splay, twist, or bend. As a result, the relevant Frank constant can be K_1, K_2, or K_3. Measuring the Frederiks threshold is often a useful technique for determining the Frank constants in the laboratory.

10.8 Defects in Nematic Phase

In the previous section, we considered the Frederiks transition as one example of a director distortion. In that example, the director distortion varies continuously as a function of the applied magnetic field: If we have a large field, we get a large director distortion; if we reduce the field to zero, the director field becomes uniform. In this section, we will consider a different type of director distortion that *cannot* just come and go continuously. These distortions are called *topological defects.*

Topological defects are large localized distortions that cannot relax away. They are an important part of liquid-crystal science (and materials science in general) for several reasons. First, the types of defects that can form in a phase depend on the symmetry of that phase. Hence, observing defects can help experimenters to recognize a phase. Second, defects determine many properties of materials, including transport properties and mechanical properties. Third, defects need to be eliminated from liquid crystals for many types of applications; understanding defects can help us to eliminate them. Finally, certain applications actually use defects, especially for bistability.

In this section, I will only describe the simplest version of defects in 2D liquid crystals. The subject of defects in 3D liquid crystals involves many subtle issues of topology, which I cannot cover here. If you are interested in that subject, I recommend the textbook *Soft Matter Physics,* by Maurice Kleman and Oleg D. Lavrentovich, as listed at the end of this chapter.

Let us begin with a 2D *polar* phase. We suppose that the magnitude of polar order is fixed, and only the direction can vary. At any point, the direction of polar order is given by a unit vector $\hat{\boldsymbol{p}}(\boldsymbol{r}) = (\cos\phi(\boldsymbol{r}), \sin\phi(\boldsymbol{r}))$, which can be represented by an arrow. Suppose this unit vector field has the configuration shown in Fig. 10.19a. This configuration is a type of topological defect, called a *vortex* or a *disclination.* We can see that this defect is a large distortion, located in the middle of the red circle, which cannot relax away. If we try to eliminate the defect by rotating some of the arrows in any local region, we will only move the defect to another position. The only way to eliminate it completely would be by moving it all the way to the edge of the material.

To characterize the defect mathematically, we use a construction called a *Burgers circuit.* Suppose we move around the red circle in the counter-clockwise direction: from east to north, west, south, and back to east. As we move around that circle, we keep track of the changes in ϕ: it increases from 0 to $\pi/2$, π, $3\pi/2$, and ends at 2π. We can write this change in ϕ as an integral $\oint d\phi = 2\pi$. If we calculate this integral about *any* loop that encloses the defect, we get the same result. The loop might be a large circle, a small circle, an ellipse, or any other shape—if the loop encloses this defect, the changes in ϕ add up to 2π. By comparison, if we calculate this integral about any loop that does not enclose a defect, the changes in ϕ add up to 0; ϕ might increase or decrease, but it will go back to its starting value. Hence, we will take the integral about the Burgers circuit as our definition of the *topological charge q* enclosed inside the loop,

$$\oint d\phi = 2\pi q_{\text{enclosed}}. \tag{10.89}$$

In particular, the example of Fig. 10.19a has a topological charge of 1. This definition of topological charge should remind you of Gauss's law in electrostatics. As in Gauss's law, we can integrate around the outer surface of a shape to find the total charge enclosed by the shape.

Next, let us consider the configuration in Fig. 10.19b. Students always guess that this configuration has a topological charge of -1. No, this guess is wrong! The analogy with electrostatics does not work that way. To determine the topological charge, let us return to the integral about the Burgers circuit. As we go around the red circle, ϕ increases from $-\pi$ to $-\pi/2$, 0, $\pi/2$, and ends at π. Hence, the net increase

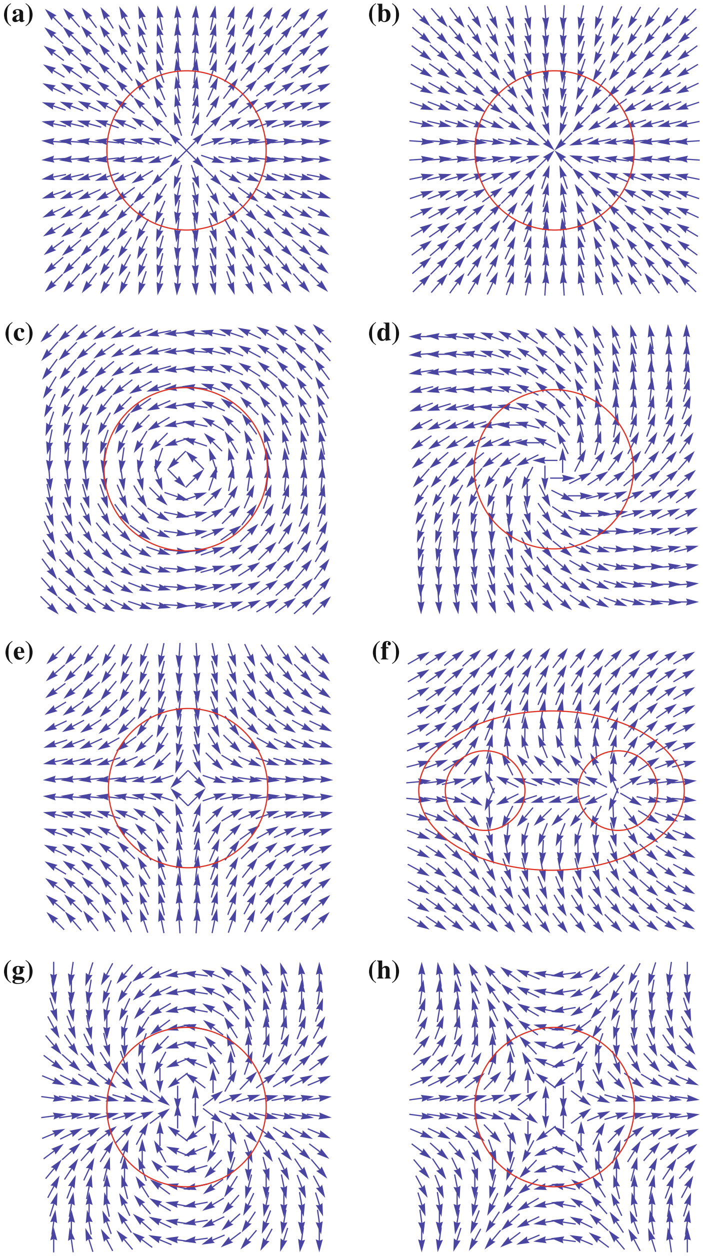


Fig. 10.19 Defects in a 2D polar phase. **a**–**d** Topological charge +1. **e** Topological charge −1. **f** Topological charges +1 and −1, combining to make 0. **g** Topological charge +2. **h** Topological charge −2 (Interactive version at http://www.springer.com/cda/content/document/cda_downloaddocument/Selinger+Interactive+Figures.zip?SGWID=0-0-45-1509169-p177545420.)

in ϕ is 2π, so $\oint d\phi = 2\pi$, just as in the previous example. Hence, this is *also* a defect with topological charge of $+1$.

We can make a similar argument for the configurations in Fig. 10.19c, d. In each of those cases, we can add up the changes in ϕ about the Burgers circuit to obtain $\oint d\phi = 2\pi$. They are also defects with topological charge of $+1$.

How can we make a defect of topological charge of -1? We need to make ϕ decrease as we go around the Burgers circuit. In other words, the polar order must rotate *clockwise* while we go around the circuit *counter-clockwise*. One of these defects is shown in Fig. 10.19e.

Now suppose we have a $+1$ defect near a -1 defect, as shown in Fig. 10.19f. If we draw a Burgers circuit about the $+1$ defect only (red circle on the right), we find that $\oint d\phi = 2\pi$. If we draw a Burgers circuit about the -1 defect only (red circle on the left), we find that $\oint d\phi = -2\pi$. If we draw a Burgers circuit about *both* of the defects (large red ellipse), we find that $\oint d\phi = 0$. Hence, the defect charges add up just as they should, based on the analogy with Gauss's law. The positive and negative charges cancel each other, when we look at a large Burgers circuit that encloses both defects. (In the interactive version of the figure, you can move the defects together and apart. If the defects come together, you can see that they annihilate like an electron and a positron, leaving a defect-free configuration.)

The polar phase can also have higher defect charges. Figure 10.19g shows a defect of topological charge $+2$, and Fig. 10.19h shows a defect of topological charge -2.

In a polar phase, the defect charge must always be an *integer,* because the polar vector field $\hat{\boldsymbol{p}}(\boldsymbol{r})$ is a single-valued function. When we go around a Burgers circuit, $\hat{\boldsymbol{p}}(\boldsymbol{r})$ must return to the same vector as at the beginning of the circuit. This is only possible if the change in $\phi(\boldsymbol{r})$ is a multiple of 2π, and hence q_{enclosed} is an integer.

At this point, let us change from the 2D polar phase to the 2D nematic phase. By analogy with the polar phase, we suppose that the magnitude of nematic order S is fixed, and only the direction can vary. At any point, the direction of nematic order is given by the director $\hat{\boldsymbol{n}}(\boldsymbol{r}) = (\cos\phi(\boldsymbol{r}), \sin\phi(\boldsymbol{r}))$. As you recall, the nematic phase has one important difference from the polar phase: the vectors $\hat{\boldsymbol{n}}$ and $-\hat{\boldsymbol{n}}$ represent exactly the same physical state. Hence, we should not illustrate the direction of nematic order by an arrow; instead, we should illustrate it by a *double-headed arrow* (or by a line segment with no arrowheads).

The nematic phase can have all of the same defects as the polar phase. For example, it can have a defect with topological charge $+1$, which might have the radial form shown in Fig. 10.20a or the tangential form shown in Fig. 10.20b. It can also have all of the other integer-charged defects discussed in the context of the polar phase.

In addition to the integer-charged defects, the nematic phase can *also* have half-integer-charged defects. For example, consider the configuration shown in Fig. 10.20c. As we move around the red Burgers circuit from east to north, west, south, and back to east, the angle ϕ increases from 0 to $\pi/4$, $\pi/2$, $3\pi/4$, and ends at π. Hence, the net increase in ϕ is π, so $\oint d\phi = \pi$. By the definition of Eq. (10.89), this defect has a topological charge of $+\frac{1}{2}$. Other half-integer charges are also possible. A $-\frac{1}{2}$ defect is shown in Fig. 10.20d, a $+\frac{3}{2}$ defect in Fig. 10.20e, and a $-\frac{3}{2}$ defect in Fig. 10.20f.

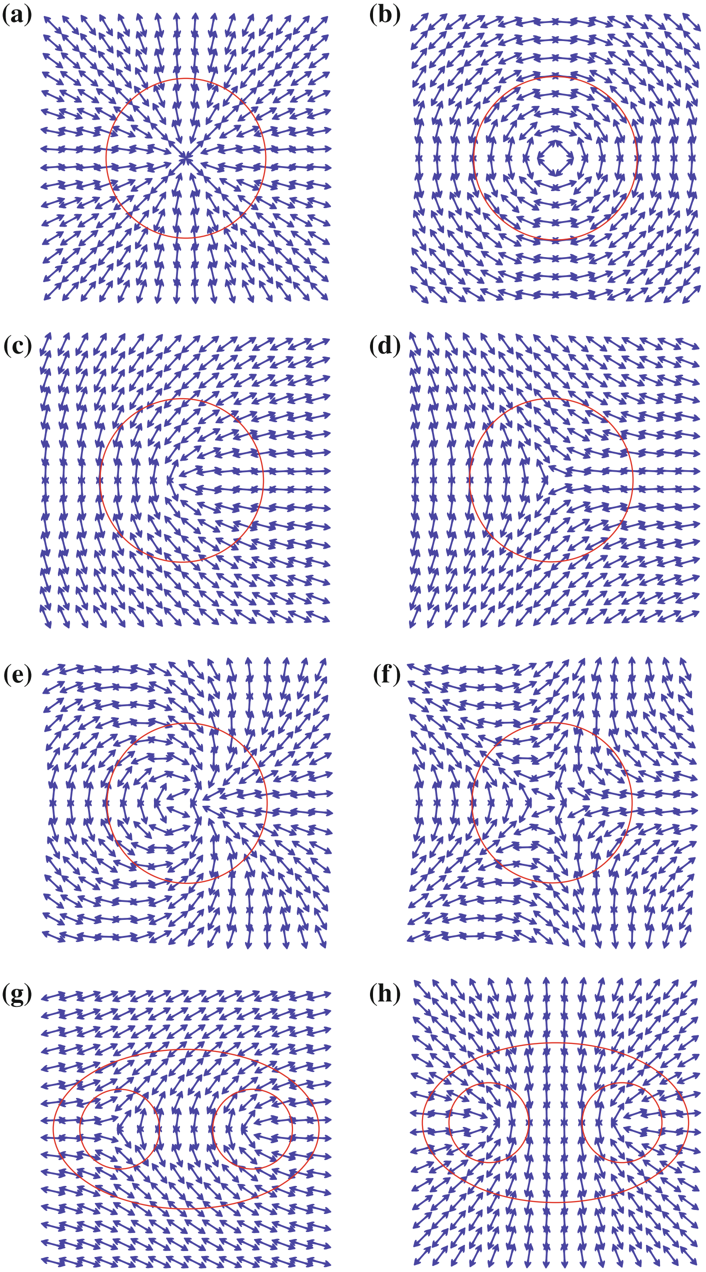


Fig. 10.20 Defects in a 2D nematic phase. **a**, **b** Topological charge $+1$. **c** Topological charge $+\frac{1}{2}$. **d** Topological charge $-\frac{1}{2}$. **e** Topological charge $+\frac{3}{2}$. **f** Topological charge $-\frac{3}{2}$. **g** Topological charges $+\frac{1}{2}$ and $-\frac{1}{2}$, combining to make 0. **h** Topological charges $+\frac{1}{2}$ and $+\frac{1}{2}$, combining to make $+1$ (Interactive version at http://www.springer.com/cda/content/document/cda_downloaddocument/Selinger+Interactive+Figures.zip?SGWID=0-0-45-1509169-p177545420.)

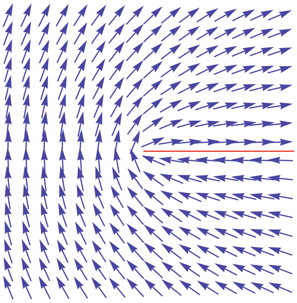


Fig. 10.21 *Unsuccessful* attempt to construct a defect of topological charge $+\frac{1}{2}$ in a polar phase. The defect is not just a point but a line with a discontinuity in the polar order parameter

The half-integer-charge defects add up appropriately, just as the integer-charge defects do. Figure 10.20g shows a $+\frac{1}{2}$ defect near a $-\frac{1}{2}$ defect. If we draw a Burgers circuit that encloses just one of these defects, we find the charge of $+\frac{1}{2}$ for the right circle and $-\frac{1}{2}$ for the left circle. If we draw a Burgers circuit that encloses both of the defects (large ellipse), we find a total topological charge of 0. Similarly, Fig. 10.20h shows two $+\frac{1}{2}$ defects near each other. Each small circle encloses a charge of $+\frac{1}{2}$, while the large ellipse encloses a total charge of $+1$.

We can see that half-integer charges are possible in the nematic phase but *not* in the polar phase. If we try to construct of defect of charge $+\frac{1}{2}$ in a polar phase, as shown in Fig. 10.21, we find that the defect is not just a point. Rather, the defect becomes a line, or *wall,* with a discontinuity in the polar order parameter. This will normally not occur in a polar phase. By contrast, there is no problem with a line defect in a nematic phase, specifically because $\hat{\boldsymbol{n}}$ and $-\hat{\boldsymbol{n}}$ represent the same physical state, so that the director field is drawn by a double-headed arrow.

Now that we have seen the *topology* of defects, let us consider the *free energy* of defects. All of the defect configurations involve substantial distortions in the director field over a long range, and hence they must cost a substantial amount of Frank free energy. To determine the free energy cost, let us use the simplest version of the Frank free energy with only a single Frank constant, from Eq. 10.72. The 2D version of this free energy is

$$F = \int d^2 r \left[\frac{1}{2} K \left(\partial_\gamma n_\alpha \right) \left(\partial_\gamma n_\alpha \right) \right]. \tag{10.90}$$

If we put our assumption $\hat{\boldsymbol{n}}(\boldsymbol{r}) = (\cos\phi(\boldsymbol{r}), \sin\phi(\boldsymbol{r}))$ into this expression, it simplifies to

$$F = \int d^2 r \left[\frac{1}{2} K |\nabla\phi|^2 \right]. \tag{10.91}$$

In a defect configuration, $\phi(\boldsymbol{r})$ is not the absolute minimum of the free energy; the absolute minimum is uniform. However, $\phi(\boldsymbol{r})$ is a *local minimum* of the free energy, because there is no small rearrangement of the director field that can reduce the free energy. Hence, it must satisfy

$$\frac{\delta F}{\delta \phi(\boldsymbol{r})} = -K\nabla^2\phi = 0. \tag{10.92}$$

This is a linear partial differential equation to characterize the director field around one or more defects. We can solve it in any coordinate system. For a single defect at the origin, it is most convenient to use polar coordinates (r, θ). For a defect of topological charge q, a solution is

$$\phi(r, \theta) = \phi_0 + q\theta. \tag{10.93}$$

This is the mathematical form that is illustrated all the pictures of single defects in Figs. 10.19, 10.20 and 10.21. In this function, when the polar coordinate θ goes from 0 to 2π, the director orientation ϕ goes from 0 to $2\pi q$, so there is a defect of charge q at the origin.

To evaluate the free energy of the defect, we calculate the gradient $\nabla\phi = (q/r)\hat{\boldsymbol{\theta}}$. We substitute that gradient back into the free energy, and set up the integral in polar coordinates, to obtain

$$F = \int_0^\infty r dr \int_0^{2\pi} d\theta \left[\frac{1}{2}K\frac{q^2}{r^2}\right] = \pi K q^2 \int_0^\infty \frac{dr}{r} = \Big[\pi K q^2 \log r\Big]_0^\infty . \tag{10.94}$$

Unfortunately, there are two problems when we try to evaluate this function at the integration limits 0 and ∞: the function $\log r$ diverges at 0 and it diverges at ∞! We need to consider each of these problems separately.

The divergence as $r \to 0$ is called an *ultraviolet divergence*. It occurs because we are trying to use our Frank elastic theory at short distances, in a regime where the theory does not apply. One problem with the theory at short distances is that it assumes that the magnitude S of nematic order is constant. This assumption is usually reasonable, because changing $\hat{\boldsymbol{n}}(\boldsymbol{r})$ only costs a little free energy, while changing S costs much more free energy. However, close to the defect, gradients in $\hat{\boldsymbol{n}}(\boldsymbol{r})$ become very large, and hence the Frank free energy for these gradients becomes huge. Hence, in a small core around the defect, the liquid crystal can melt into the isotropic phase, with $S \to 0$, so that it does not need to pay the high free energy cost for director variations. We might model this defect core melting by going back to a theory in terms of the nematic order tensor $Q_{\alpha\beta}(\boldsymbol{r})$. More commonly, people just assume that there is some core radius a, of order the intermolecular spacing of nanometers, and use the Frank elastic theory only for lengths greater than this core radius. Inside the core, they just assume there is an energy E_{core}.

The divergence as $r \to \infty$ is called an *infrared divergence*. It is a real physical divergence, because it occurs in the long length scale where Frank elastic theory really does apply. However, it is less serious than you might fear. Physically, we need to cut off the integral at a length scale corresponding to the size R_{max} of the sample. As a result, the free energy of the defect becomes

$$F = E_{\text{core}} + \pi K q^2 \int_a^{R_{\max}} \frac{dr}{r} = E_{\text{core}} + \pi K q^2 \log\left(\frac{R_{\max}}{a}\right). \tag{10.95}$$

The logarithm diverges in the limit of infinite system size, but it diverges very slowly! For example, suppose we have a system size $R_{\max} = 1\,\text{cm}$ and a core radius $a = 1\,\text{nm}$. The logarithm would just be a very modest factor of $\log 10^7 \approx 16$. If $R_{\max} = 1\,\text{m}$, the logarithm would be $\log 10^9 \approx 21$. If $R_{\max} = 1\,\text{km}$, it would be $\log 10^{12} \approx 28$, which is still reasonable. Hence, we do not really need to worry about this factor.

The most interesting part of the defect free energy is the factor of q^2. It shows that defects with high topological charge have a much greater free energy than defects with low charge. Hence, there is an energetic preference for high-charge defects to break up into defects with the minimum possible charge (which is 1 for a polar phase or $\frac{1}{2}$ for a nematic phase). For example, if a defect of charge 1 breaks into two defects of charge $\frac{1}{2}$, the energy decreases from $(1)^2$ to $2 \times (\frac{1}{2})^2$. Hence, in experiments, we should expect to see the lowest-charge defects but not higher-charge defects.

We can use a similar method to calculate the interaction between two defects. Suppose we have a defect of charge q_1 at $(x_0, 0)$, and another defect of charge q_2 at $(-x_0, 0)$. Because the Euler-Lagrange equation (10.92) is linear, the solution for $\phi(x)$ is the superposition of the solutions for each individual defect. In Cartesian coordinates, it can be written as

$$\begin{aligned}\phi(x, y) &= \phi_1(x, y) + \phi_2(x, y) \qquad (10.96)\\ &= \phi_1 + q_1 \tan^{-1}\left(\frac{y}{x - x_0}\right) + \phi_2 + q_2 \tan^{-1}\left(\frac{y}{x + x_0}\right).\end{aligned}$$

Hence, the extra free energy associated with the interacting pair (compared with the free energies of isolated defects 1 and 2) is

$$\begin{aligned}F_{\text{int}} &= F_{\text{pair}} - F_1 - F_2\\ &= \frac{1}{2} K \int_{-X_{\max}}^{X_{\max}} dx \int_{\infty}^{\infty} dy \left[|\nabla\phi|^2 - |\nabla\phi_1|^2 - |\nabla\phi_2|^2\right]\\ &= \frac{1}{2} K \int_{-X_{\max}}^{X_{\max}} dx \int_{\infty}^{\infty} dy \frac{2q_1 q_2 (x^2 - x_0^2 + y^2)}{(x^2 - 2xx_0 + x_0^2 + y^2)(x^2 + 2xx_0 + x_0^2 + y^2)}\\ &= 2\pi K q_1 q_2 \log\left(\frac{X_{\max}}{x_0}\right). \qquad (10.97)\end{aligned}$$

From this result, we see that the interaction free energy is proportional to the product of charges $q_1 q_2$, just as in Coulomb's law for electrostatics. The interaction free energy scales as the logarithm of the separation between defects, and hence the force scales as 1/separation, which is analogous to Coulomb's law in two dimensions. The force is repulsive for defects of the same sign, and attractive for defects of opposite signs.

10.9 Chirality and Cholesteric Liquid Crystals

So far we have assumed that the liquid-crystal molecules can be represented as rods or arrows. In many cases, this is a reasonable approximation. However, there is one type of molecular asymmetry that has a profound effect on the large-shape structure of a liquid-crystal phase. This asymmetry is called *chirality.*

The concept of chirality refers to an object that is different from its mirror image; it cannot be superimposed on its mirror image through any rotations. Of geometrical shapes, a sphere or a cylinder is not chiral, because it is equivalent to its mirror image. However, a hand is chiral, because a right hand is different from its mirror image, which is a left hand. Hence, we can say that a chiral object has *handedness.*

Of course, there are many ways to form chiral objects. In chemistry, the most common way to form chiral molecules is through chemical bonds on carbon atoms. As you may recall from chemistry classes, a carbon atom normally bonds to four atoms or chemical groups, which are arranged in a tetrahedron around the central carbon. If at least two of these four chemical groups are equivalent to each other, then the molecule is equivalent to its mirror image, and it is called *achiral.* However, if all four of these chemical groups are different from each other, then the molecule becomes different from its mirror image, and is called *chiral.*

An example is the chemical structure of alanine, shown in Fig. 10.22. Here, the central carbon is bonded to H, CH_3, NH_2, and COOH. Because these chemical groups are all different from each other, the molecule is not equivalent to its mirror image; there is no way to rotate the molecule so that it can be superimposed on its mirror image. The molecule can occur in two distinct mirror-image forms, called *stereoisomers.* There are several chemical conventions for how to label these forms; in one convention they are called *(R)*- and *(S)*-alanine (after the Latin words for right and left).

Synthesizing chiral molecules is an important part of organic chemistry, mainly because of biomedical applications. Because the human body is composed of chiral molecules (including proteins, lipids, and DNA), different stereoisomers of drugs interact with the body in different ways. One well-known example is thalidomide: one form of the drug is useful for preventing nausea in pregnant women, while the mirror image is infamous for causing birth defects. Hence, the pharmaceutical industry puts great effort into developing chirally pure drugs.

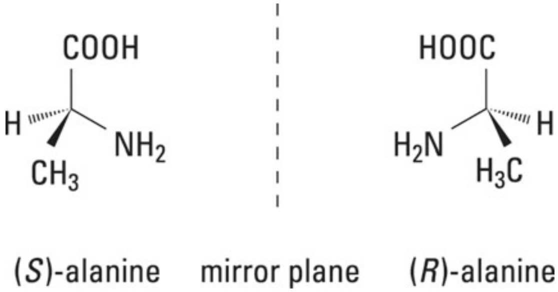


Fig. 10.22 Structure of alanine, a typical chiral molecule, showing the difference between the two mirror images. From http://www.nobelprize.org/nobel_prizes/chemistry/laureates/2001/popular.html

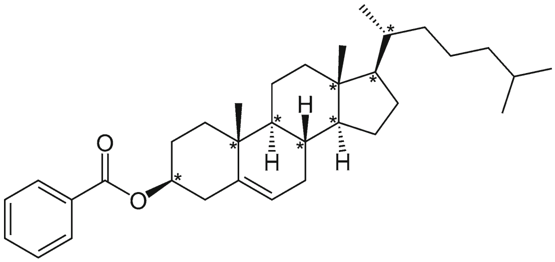


Fig. 10.23 Chemical structure of the liquid crystal cholesteryl benzoate, with stars indicating the chiral carbons. Adapted from http://en.wikipedia.org/wiki/Cholesteryl_benzoate

Like drugs, some types of liquid-crystal molecules can also exist in distinct chiral forms, which can be formed through biological processes or chemical synthesis. As an example, Fig. 10.23 shows the chemical structure of cholesteryl benzoate, which was the first material in which a liquid-crystal phase was discovered (by the Austrian botanist Friedrich Reinitzer). This molecule includes eight chiral carbons, which are indicated by stars in the figure. Because each chiral carbon can have take the *(R)* or *(S)* form, the molecule can in principle have $2^8 = 256$ distinct stereoisomers, only one of which occurs naturally. In this respect, cholesteryl benzoate is quite different from 5CB, shown in Fig. 10.1a. The liquid crystal 5CB does not have any chiral carbons; it has a symmetric structure that is equivalent to its mirror image.

Now the question is: If we form a liquid-crystal phase of asymmetric molecules like cholesteryl benzoate, how does the molecular chirality affect the structure of the phase? We can think about this question either microscopically or macroscopically.

On a microscopic level, we can visualize the packing of chiral molecules by the packing of hard screws, as shown in Fig. 10.24. Screws are chiral objects, which are different from their mirror images. When screws pack in the most efficient possible way, so that their threads fit together, they are not parallel to their neighbors. Rather,

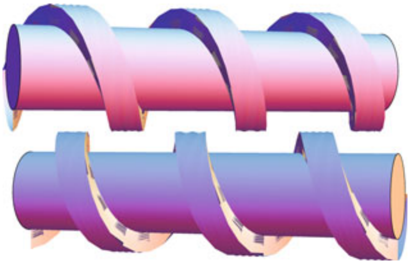


Fig. 10.24 Schematic representation of the packing of chiral molecules by the packing of hard screws. In the most efficient packing, molecules are not parallel to their neighbors, but rather are at a slight twist angle with respect to their neighbors. Based on J.P. Straley, *Phys. Rev. A* **14**, 1835 (1976) (Interactive version at http://www.springer.com/cda/content/document/cda_downloaddocument/Selinger+Interactive+Figures.zip?SGWID=0-0-45-1509169-p177545420.)

they are at a slight twist angle with respect to their neighbors. Of course, most chiral molecules do not look very much like screws, and we cannot take this packing model too seriously. However, it illustrates a point of symmetry: Chiral molecules pack together with a certain favored twist, and the mirror-image molecules pack together with the opposite favored twist.

On a macroscopic level, we can see the same point by thinking about order and symmetry. A chiral material is less symmetric than an achiral material, because the achiral material has a reflection symmetry but the chiral material does not. Likewise, a chiral material has more order than an achiral material, because the chiral material has a particular handedness. As a result, the free energy for a chiral liquid crystal can have an extra term, which is not permitted in the free energy for an achiral liquid crystal.

In an achiral liquid crystal, the Frank free energy of Eq. (10.73) has terms that are *quadratic* in splay, twist, and bend. In a chiral liquid crystal, the Frank free energy can also have an extra term that is *linear* in twist. This extra term is $-K_2 q[\hat{\boldsymbol{n}} \cdot (\nabla \times \hat{\boldsymbol{n}})]$; the coefficient is written as $K_2 q$ for a reason that will become clear soon. This term breaks the symmetry between right-handed twist $[\hat{\boldsymbol{n}} \cdot (\nabla \times \hat{\boldsymbol{n}}) > 0]$ and left-handed twist $[\hat{\boldsymbol{n}} \cdot (\nabla \times \hat{\boldsymbol{n}}) < 0]$. If $K_2 q > 0$, then right-handed twist is favored and left-handed twist is disfavored; the opposite is true if $K_2 q < 0$. Mathematically, we can say that the twist $\hat{\boldsymbol{n}} \cdot (\nabla \times \hat{\boldsymbol{n}})$ is a pseudoscalar, and the coefficient $K_2 q$ is also a pseudoscalar; their product is a proper scalar, which is permitted in the free energy. The pseudoscalar coefficient $K_2 q$ is only permitted in a chiral liquid crystal; it must be zero in an achiral liquid crystal because of the reflection symmetry.

With the extra term, the Frank free energy of a chiral liquid crystal becomes

$$F = \int d^3 r \left[\frac{1}{2} K_1 (\nabla \cdot \hat{\boldsymbol{n}})^2 + \frac{1}{2} K_2 [\hat{\boldsymbol{n}} \cdot (\nabla \times \hat{\boldsymbol{n}})]^2 + \frac{1}{2} K_3 |\hat{\boldsymbol{n}} \times (\nabla \times \hat{\boldsymbol{n}})|^2 \right. \\ \left. - K_2 q [\hat{\boldsymbol{n}} \cdot (\nabla \times \hat{\boldsymbol{n}})] \right]. \tag{10.98}$$

By completing the square, we can transform it into

$$F = \int d^3 r \left[\frac{1}{2} K_1 (\nabla \cdot \hat{\boldsymbol{n}})^2 + \frac{1}{2} K_2 [\hat{\boldsymbol{n}} \cdot (\nabla \times \hat{\boldsymbol{n}}) - q]^2 + \frac{1}{2} K_3 |\hat{\boldsymbol{n}} \times (\nabla \times \hat{\boldsymbol{n}})|^2 + \text{const} \right]. \tag{10.99}$$

Now you see why the coefficient of the linear term was written as $K_2 q$: to make it easier to complete the square. From Eq. (10.99), we can see that the minimum free energy for a chiral liquid crystal has zero splay, nonzero twist, and zero bend:

$$\nabla \cdot \hat{\boldsymbol{n}} = 0, \tag{10.100a}$$

$$\hat{\boldsymbol{n}} \cdot (\nabla \times \hat{\boldsymbol{n}}) = q, \tag{10.100b}$$

$$\hat{\boldsymbol{n}} \times (\nabla \times \hat{\boldsymbol{n}}) = 0. \tag{10.100c}$$

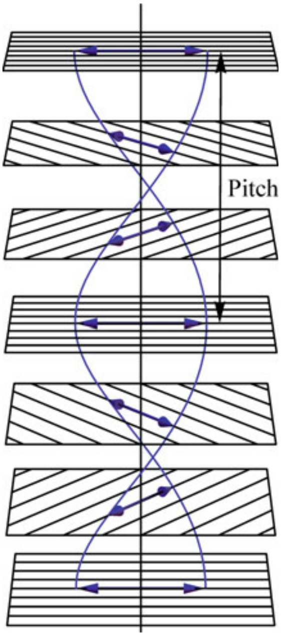


Fig. 10.25 Director configuration in a chiral nematic or cholesteric phase (Interactive version at http://www.springer.com/cda/content/document/cda_downloaddocument/Selinger+Interactive+Figures.zip?SGWID=0-0-45-1509169-p177545420.)

Is it possible for the liquid crystal to find any director configuration that satisfies the three equations (10.100)? Yes! The solution is the helical structure shown in Fig. 10.25. Mathematically, the director field can be written as

$$\hat{\boldsymbol{n}}(\boldsymbol{r}) = (\cos qz, \sin qz, 0). \tag{10.101}$$

This structure is called the *chiral nematic* or *cholesteric* phase. In the cholesteric phase, the local structure is similar to the nematic phase: the molecules are partially aligned along the local axis $\hat{\boldsymbol{n}}(\boldsymbol{r})$. However, on a larger scale, the axis $\hat{\boldsymbol{n}}(\boldsymbol{r})$ varies as a function of position in this helical configuration.

Along the helical axis, the cholesteric phase is periodic, and the periodicity is called the *pitch*. In terms of the parameter q, the pitch is π/q. (It is π/q, not $2\pi/q$, because $\hat{\boldsymbol{n}}$ and $-\hat{\boldsymbol{n}}$ describe the same physical state.) In typical cholesteric liquid crystals, the pitch is on the order of microns, similar to the wavelength of visible light or the size of a biological cell. It is three orders of magnitude bigger than the spacing between molecules, which is on the order of nanometers. Hence, the average angle between neighboring moleculees is of order $\pi/1000$ radians, which is much smaller than we might guess from the packing of screws in Fig. 10.24. The favored twist from molecule to molecule is a very small effect, but it adds up in a system of many chiral molecules.

The cholesteric pitch generally depends on temperature. Because of this dependence, one application of cholesteric liquid crystals is for a thermometer, which changes its pitch (and hence its color) as a function of temperature.

10.10 Other Liquid-Crystal Phases

In addition to the nematic and cholesteric phases, there is a wide variety of other liquid-crystal phases, with different types of order that are intermediate between isotropic liquids and fully ordered crystals. Here, I can just mention a few of the highlights.

One class of liquid-crystal phases is called *blue phases*. Like the cholesteric phase, blue phases occur in systems of chiral liquid crystals, and the molecular chirality leads to a spontaneous twist of the director field. However, while the cholesteric phase has a simple twist of the director in one direction, blue phases have a more complex 3D modulation of the director field. This modulation can be regarded as a lattice of double-twist tubes, separated by a network of defect lines.

Another class of liquid-crystal phases is called *smectic phases*. In smectic phases, the molecules have nematic orientational order and they are arranged in layers, i.e., there is a 1D density wave. Hence, a smectic phase is a crystal in one direction and a fluid in the other two directions.

Because smectic phases have *both* nematic order and smectic layer order, we need to specify the relationship between these two types of order. If the nematic order is aligned perpendicular to the smectic layers (so that the molecules stand upright with respect to the layers), then the liquid crystal is called a *smectic-A* phase. If the nematic order is aligned at an oblique angle to the smectic layers (so that the molecules are tilted with respect to the layers), then the liquid crystal is called a *smectic-C* phase. Comparing these two phases, the smectic-A phase is symmetric under arbitrary rotations about the layer normal, while the smectic-C phase is not symmetric under such rotations. Hence, the smectic-C phase has less symmetry and more order than the smectic-A phase.

Apart from molecular tilt, smectic liquid crystals can have other types of order within the layers. For example, they can have *hexatic* order in the orientations of the vectors between neighboring molecules within the layers. (These vectors are called *bonds,* although they are not chemical bonds, and hence the order is called *bond-orientational order.*) If the liquid crystal has both hexatic order and tilt order, then it can have different phases with different relative orientations of these two types of order. Combining these various types of order, there can be a wide range of smectic phases.

If a smectic-C phase is made of chiral molecules, the combination of smectic order *and* tilt order *and* chirality leads to a *ferroelectric liquid crystal,* with a spontaneous electrostatic polarization. This electrostatic polarization is a vector in the smectic layer plane, perpendicular to the director. Hence, this phase really does have polar order, in contrast with the tensor order in a nematic. In a bulk ferroelectric liquid crystal, the tilt direction and the polarization tend to rotate from layer to layer in a helix, analogous to a cholesteric phase. Near a surface, the helix can be suppressed by surface interactions, leading to a *surface-stabilized ferroelectric liquid crystal.*

In some cases, a smectic liquid crystal of chiral molecules can form a *twist-grain-boundary (TGB) phase,* shown in Fig. 10.26. This structure consists of a series

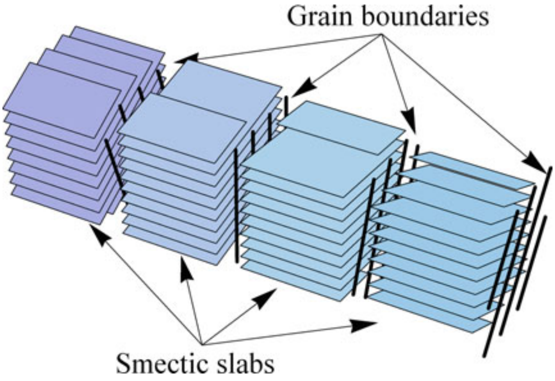


Fig. 10.26 Structure of the twist-grain-boundary (TGB) phase (Interactive version at http://www.springer.com/cda/content/document/cda_downloaddocument/Selinger+Interactive+Figures.zip?SGWID=0-0-45-1509169-p177545420.)

of smectic slabs, separated by grain boundaries. Across each grain boundary, the orientation of the smectic slabs rotates around a helical axis, with a twist induced by the molecular chirality. If one looks closely at the grain boundaries, one can see that each grain boundary is composed of a series of line defects in the smectic order, which are called *screw dislocations.* Although this theoretical construction is quite complex, the TGB phase has been seen experimentally in X-ray diffraction and freeze-fracture electron microscopy.

In research on all of these liquid-crystal phases, there are several common themes. First is the concept of order as broken symmetry. All of these phases have different types of order, meaning that they have broken different types of symmetry, compared with a uniform, isotropic state. Each type of broken symmetry is characterized by an order parameter, which indicates the magnitude (how much the symmetry is broken) and direction (in which way the symmetry is broken). Each phase can be described by an order parameter, or by the relationships among multiple order parameters.

A second general theme is that minimization of the free energy determines the equilibrium state of the material. In most cases, this minimization involves a competition between energy (which favors an ordered phase) and entropy (which favors a disordered phases). In certain systems, however, the transition between order and disorder is driven by competition between two kinds of energy or two kinds of entropy.

Finally, another general theme is the concept that ordered phases have topological defects. In general, each type of order is associated with its own type of defect. These defects help us to recognize a phase, and often are very important for the properties of the phase.

In any area of research on soft materials, these concepts will take you far!

Further Reading

Introductory, non-technical books:

1. P.J. Collings, *Liquid Crystals: Nature's Delicate Phase of Matter, Second Edition* (Princeton, 2001)

2. M.R. Fisch, *Liquid Crystals, Laptops And Life* (World Scientific, 2004)

Intermediate level:

3. P.J. Collings, M. Hird, *Introduction to Liquid Crystals: Chemistry and Physics* (Taylor & Francis, 1997)
4. A. Jákli, A. Saupe, *One- and Two-Dimensional Fluids: Properties of Smectic, Lamellar and Columnar Liquid Crystals* (Taylor & Francis, 2006)

Advanced level:

5. P.G. de Gennes, J. Prost, *The Physics of Liquid Crystals, Second Edition* (Oxford, 1995)
6. P.M. Chaikin, T.C. Lubensky, *Principles of Condensed Matter Physics* (Cambridge, 1995)

Special emphasis on defects, as well as other topics:

7. M. Kleman, O.D. Lavrentovich, *Soft Matter Physics: An Introduction* (Springer, 2003)

Historical perspective:

8. D. Dunmur, T. Sluckin, *Soap, Science, and Flat-Screen TVs: A History of Liquid Crystals* (Oxford, 2011)

Index

J.V. Selinger, *Introduction to the Theory of Soft Matter*,
Soft and Biological Matter, DOI 10.1007/978-3-319-21054-4