CONCEPTS
IN STATISTICAL MECHANICS

ARTHUR HOBSON

To Dodie, Lisa and Ben

Preface

IT HAS ALWAYS SEEMED to the author that science is the attempt to *understand* nature, and that "understanding nature" means something deeper than merely predicting the results of experiments; to "understand" means to explain in terms of a few clearly stated underlying concepts. This is a book about the concepts underlying statistical mechanics, and about the explanation of certain physical laws in terms of those concepts. The book emphasizes the conceptual problems and the underlying mathematical structure of statistical mechanics, rather than applications. Most books on statistical mechanics give an adequate treatment of techniques and applications; the reader is referred to these books for information on the calculation of partition functions, specific heats, etc. It is hoped that the reader of the present book will come away with some understanding of why these calculations work.

The central object in the theory of statistical mechanics is the N-body distribution function $\varrho(x, t)$, where x represents the "phase point" for the entire N-body system. Thus, any coherent theory of statistical mechanics must give a clear answer to the question "What is the meaning of ϱ?" As discussed in Chapter 1, there are several widely held answers to this question; the different answers actually amount to different points of view toward the scope of statistical mechanics. Statistical mechanics is sometimes regarded as the study of macroscopic systems in terms of their microscopic properties, sometimes as the study of time averages of mechanical quantities, and sometimes as the study of incompletely specified mechanical systems. The point of view adopted in this book is that statistical mechanics is the study of incompletely specified mechanical systems; as discussed in Chapter 1, this point of view appears to have greater clarity and usefulness than the other points of view.

Having adopted, in Chapter 1, the view that statistical mechanics is the study of mechanical systems in terms of incomplete information, we are led to regard ϱ as a probability distribution representing the information which we *do* have about the system. Probability theory thus becomes the basic mathematical language of statistical mechanics; this language is studied in Chapter 2.

The branch of probability theory known as *information theory* deals with the measurement, acquisition, and transmission of information. If one believes that statistical mechanics is the study of systems in terms of incomplete information, then one is immediately led to try applying information theory to statistical mechanics. E. T. Jaynes (see Ref. 4 of Chapter 1) was the first to establish a clear and useful connection between information theory and statistical mechanics, although several authors prior to Jaynes speculated on qualitative connections between the two fields (see e.g. Refs. 15, 16, and 17 of Chapter 1). This book is based on the information theory approach to statistical mechanics. The relevant portions of information theory are presented in Chapter 2.

Chapters 3 and 4 deal with the fundamental concepts of classical and quantum statistical mechanics. Topics discussed include: Liouville's equation, statistical correlations, the significance of large numbers of particles in statistical mechanics, the choice of an appropriate initial distribution, equilibrium, the laws of equilibrium thermodynamics, thermodynamic entropy, and the justification of the methods of statistical mechanics. Most of the mathematical formalism of these chapters will be familiar to anyone who has had a course in statistical mechanics; however, it is hoped that the reader will obtain a deeper understanding of the ideas behind the formalism.

Chapter 5 deals with the fascinating and much-debated topic of irreversibility. Topics discussed include: irreversibility in classical and quantum statistics, the definition of a non-equilibrium entropy, and the generalization (to non-equilibrium situations) of the second law of thermodynamics.

This book should be accessible to anyone who has learned classical mechanics (preferably at the level of H. Goldstein's *Classical Mechanics*), quantum mechanics (preferably including Dirac notation), and thermodynamics. Previous background in statistical mechanics is not essential, although the reader might find such background helpful. The mathematical background acquired by physics undergraduates should be adequate.

The book should be of value to scientists and students who desire a broad understanding of statistical mechanics without being burdened by calculational details, as well as to researchers in statistical mechanics who desire deeper insight into the foundations. Although this book is not a textbook, it may be useful as a supplementary text in statistical mechanics since most existing textbooks deal mainly with techniques and results and treat the concepts only superficially. Philosophers and scientists interested in the role of probability theory and inductive reasoning in science, or in the philosophical significance of irreversibility and the second law, may find the book helpful.

ACKNOWLEDGMENTS

I would like to thank the National Science Foundation for providing a research grant while this book was written, and the Aspen Center For Physics for providing a beautiful and stimulating environment during the summer of 1969. I also thank my many friends whose comments helped clarify fuzzy points. In particular, I would like to thank Stephen Schery and Stephen Day for useful discussions and comments during the writing of this book, and Baldwin Robertson for a valuable and creative review of the manuscript.

University of Arkansas

ARTHUR HOBSON

Contents

CHAPTER 1

What is Statistical Mechanics

SUPPOSE YOU MUST PREDICT the behavior of three billiard balls, given only the approximate initial velocity of the cue ball. Or suppose you need to calculate the radiative properties of a hydrogen atom located in an imprecisely known electromagnetic field. Again, you might be asked to predict the evolution of a mole of gas, given that at $t = 0$ all the molecules were on the left hand side of the container. Or perhaps you wish to investigate the motion of a harmonic oscillator, given only the approximate value of the spring constant.

What do these problems have in common? Well, for one thing, each deals with a fairly common classical or quantum mechanical system. For another thing, these problems do not appear in textbooks or courses on mechanics. Mechanics deals with *precisely specified* problems: precisely specified initial conditions, precisely specified electromagnetic fields, precisely specified spring constants. The above problems are not of this sort; the complete information required for a mechanical treatment is not given. Yet scientists must often solve just such problems.

What should you do with such a problem? Should you just throw up your hands and pronounce the problem unsolvable in the absence of complete information, or would it perhaps be better to use the information which *is* given in an attempt to make the most reasonable predictions as to what is likely to happen to the system? That is, would it be better to *guess*? The answer, obviously, is that you should guess. The scheme which has been worked out for making such guesses is called *statistical mechanics*. Mechanics, because it deals with mechanical systems. Statistical, because it deals with *imprecisely specified* (or *statistically specified*) systems, and makes only imprecise (or statistical) predictions.

The theory of mechanics (whether classical or quantum) applies only to the idealized situation in which the precise initial state is known, and all internal and external influences on the evolution of the system are precisely known. Now, the fact that mechanics is an idealization should not bother us unduly, since all physical theories are idealizations of one sort or another. But if problems arise which do not conform reasonably well to the idealization involved in a particular theory, then one looks around for a different

theory. Statistical mechanics is the theory within which the idealized assumption of complete information is removed.

Let us make these ideas more precise. The **state of the system** will mean a complete description of a system at one instant of time, i.e. a description which is as complete and detailed as is permitted by the laws of physics. In classical mechanics, the state is determined by measuring the coordinates and momenta of all the degrees of freedom, and is represented by a point $(q_1, q_2, \ldots, q_f, p_1, p_2, \ldots, p_f)$ in a $2f$-dimensional phase space, where f is the number of degrees of freedom; in quantum mechanics, the state is determined by making precise measurements of a complete set of commuting observables, and is represented by a function $\Psi(q_1, q_2, \ldots, q_f)$. Any measurement which determines the precise state of a quantum or classical mechanical system will be called a **complete measurement**.

In this book, we will deal only with non-relativistic Hamiltonian systems. That is, classical systems are assumed to possess a Hamiltonian function $H(q_1, \ldots, q_f, p_1, \ldots, p_f, t)$ and to be governed by Hamilton's equations; quantum systems are assumed to possess a Hamiltonian operator $\hat{H}(t)$ and to be governed by Schroedinger's equation.

Mechanics is the study of the time evolution of the state of a system. It is a scheme for predicting the precise state at t_1, provided only that the precise state is known at some other time t_0 and that the Hamiltonian is known between t_0 and t_1. Thus, mechanics (whether classical or quantum) allows us to predict *with certainty* the state at t_1, in terms of a complete measurement at t_0 and the Hamiltonian between t_0 and t_1. Note that only in the case of classical mechanics does this scheme allow us to predict with certainty the outcome of all experiments performed on the system at t_1, since for quantum systems even a complete description does not provide a prediction with certainty of the outcome of every possible experiment.

Briefly, *mechanics is the study of completely specified systems.*

Statistical mechanics is the study of mechanical systems in situations where the description is incomplete. It is a scheme for making reasonable predictions about the outcome of experiments at t_1, in terms of an incompletely specified state at t_0 and/or an incompletely specified Hamiltonian between t_0 and t_1. Such predictions cannot be made with certainty; all that statistical mechanics will do is provide reasonable guesses.

Briefly *statistical mechanics is the study of incompletely specified systems.*

The viewpoint that statistical mechanics is the study of incompletely specified mechanical systems has been stated explicitly by R.C.Tolman[1] and others[2-4], and (as will be discussed below) appears to be implicit in the

work of J.W.Gibbs. There are, however, at least two other widely held points of view toward the scope of statistical mechanics; we will refer to them as the **macroscopic viewpoint** and the **ergodic viewpoint**. Since this is a book about concepts, and since a first step in establishing concepts should be to agree upon what subject it is we are discussing, it is appropriate to consider the different points of view toward the scope of statistical mechanics.

The **macroscopic viewpoint** holds that statistical mechanics is the study of macroscopic systems in terms of their microscopic properties. Put another way, this viewpoint holds that statistical mechanics is the study of systems having a large number of degrees of freedom, or even that statistical mechanics is only valid in the limiting (idealized) case of an infinite number of degrees of freedom. For example, Grad[5] states that "the single feature which distinguishes statistical mechanics from mechanics is the large number of degrees of freedom". Thus, in this viewpoint, statistical mechanics is merely the mechanics of many-body systems. This view has much to recommend it: most applications of statistical mechanics have been to large systems; this viewpoint is easy to grasp; it does not involve the somewhat subjective notion of a "lack of complete information". However, the following points should be noted:

1) Systems occurring in nature have a finite number of degrees of freedom.* The idealization to an infinite number of degrees of freedom is sometimes useful, but situations may arise in which the consequences of a finite number of degrees of freedom need to be retained without, however, going over to a purely mechanical treatment. For example, the statistical mechanical explanations of many phenomena (such as the scattering of light from density fluctuations in the atmosphere) require an analysis of the statistical fluctuation, or variance (see Section 2.1), of some physical observable; such phenomena require the consideration of finite systems, since statistical fluctuations usually vanish in the limit $N \to \infty$.

2) Problems such as the billiard ball, hydrogen atom, and the harmonic oscillator examples described in the opening paragraph may actually arise. If one is willing to apply the formalism of statistical mechanics only to large systems, then one is unable to make any predictions for such problems.

3) As will be seen in the remainder of this book, at no point in the general formalism of statistical mechanics is the number of degrees of freedom

* Continuous fields, when treated as mechanical systems, are an exception to this statement. For instance, an electromagnetic field may be thought of as a mechanical system having an infinite number of degrees of freedom.

required to be large. Thus, from a purely formal point of view, the restriction to large systems appears superfluous.

4) In theories taking the macroscopic viewpoint but *not* the ergodic viewpoint, the precise meaning of the distribution function ϱ (to be introduced in Chapter 3 for classical systems, and Chapter 4 for quantum systems) is generally quite obscure. Is ϱ a probability distribution, and if so what is the origin of the probability? How is ϱ to be determined in specific situations? Does ϱ somehow represent the exact mechanical state?

The **ergodic viewpoint** holds that statistical mechanics is the study of the infinite time averages of mechanical observables. For classical systems, the time average of a phase function $g(x)$ is defined by

$$\overline{g(x)} = \lim_{T\to\infty} \frac{1}{T} \int_0^T g\{x(t)\}\, dt,$$

where x is the $2f$-dimensional phase point, and $x(t)$ represents the mechanical motion. An **ergodic system** is a system for which time averages are equal to the standard "ensemble" averages of statistical mechanics (see Chapter 3), so that time averages may be calculated, without finding the exact motion of the system, from the standard distribution function formalism of statistical mechanics. For example, Truesdale[6] states that "the purpose of statistical mechanics, for purposes of equilibrium, is to calculate time averages, and the ensemble theory is useful only as a tool enabling us to calculate time averages without knowing how to integrate the equations of motion". This viewpoint has often been associated with the macroscopic viewpoint, since it has historically been supposed that only large systems are ergodic. However, recent results[7] indicate that the size of the system is in no way related to ergodicity.

Again, the ergodic viewpoint has much to recommend it: macroscopic measurements are never instantaneous, so that measured values are always time averages (but not necessarily *infinite* time averages); the meaning of the distribution function ϱ is clear in this viewpoint (ϱ is merely a mathematical device to help in the calculation of time averages); this point of view has led to many useful and beautiful new results in mathematics and in the general theory of mechanical systems. However, the following points should be noted:

1) Real measurements are made over *finite* time intervals. Infinite time averages *never* apply to non-equilibrium (i.e. time-dependent) phenomena: it is obvious from their definition that infinite time averages cannot themselves depend on t. Thus (as is indeed stated above by Truesdale), the ergodic viewpoint restricts statistical mechanics to the study of equilibrium phenomena. But the general formalism of statistical mechanics applies perfectly well to non-equilibrium situations. Restricting statistical mechanics to equilibrium phenomena is analogous to applying Maxwell's equations only to static fields or applying Newton's laws only to stationary systems!

2) Even within the framework of equilibrium phenomena, the theorems (known as *ergodic theorems*) necessary to vindicate the ergodic viewpoint have yet to be established. Briefly, an ergodic theorem means any theorem which establishes, or partially establishes, the equality of infinite time averages and certain "ensemble averages" (see Chapter 3). It seems that the only system for which the necessary theorems have been established is the N-body hard sphere gas in a perfectly reflecting box[7], and even in this case most physicists must accept the result mainly on faith since the proof involves a sophisticated use of measure theory and differential geometry. It should be noted that Sinai[7] has proved that any hard sphere gas of two or more particles is ergodic, so there seems to be no connection between large systems and ergodicity.

3) Even if one restricts statistical mechanics to equilibrium phenomena, and assumes that all physically reasonable systems are ergodic, one must still add probability assumptions (of a fairly weak variety) to the theory in order to apply the theory to the real world. The reason for this is quite simple: even if the system is ergodic, there are still certain exceptional mechanical motions in phase space along which time averages and phase averages are not equal; i.e. ergodic systems exhibit the required equality not on *all*, but only on *almost all*, orbits. Thus, one must assume that the exceptional orbits are *highly improbable*. So even if the ergodic theorems are true, they cannot be used to reduce statistical mechanics to pure mechanics[8,9].

Having considered the macroscopic and ergodic views, we will now discuss the viewpoint taken in this book, namely that statistical mechanics is the study of incompletely specified mechanical systems.

1) This viewpoint includes the macroscopic view as a special case, since large systems are always incompletely specified. For example, it would take an army of graduate students working over a time-span of something like the age of the universe just to write down the mechanical initial conditions for each particle in a mole of gas. The cost of the paper involved would no doubt

exceed even the U.S. national debt! Thus we never possess complete information about the initial state of a macroscopic system. Furthermore, even if we did possess complete initial data, we would not want to use it because the exact equations of motion would be too cumbersome to solve and because the resulting predictions would be far too detailed to be of any value. Thus, we would choose to throw out some of the data in order to simplify the description.

2) Since data about time averages are a form of incomplete information, this viewpoint also includes the ergodic viewpoint as a special case, but without requiring the proof of any ergodic theorem.

3) This viewpoint applies to any mechanical system, including small ones, for which the given information is less than complete. This viewpoint represents the broadest possible interpretation of the formalism of statistical mechanics.

4) In the author's opinion, the viewpoint taken in this book is pedagogically superior to both the macroscopic viewpoint (which is conceptually unclear, especially as regards the meaning of ϱ) and the ergodic viewpoint (which is mathematically difficult). It allows one to treat simple, one-body and two-body systems within the framework of statistical mechanics; such treatments have obvious pedagogical advantages. This view, especially as it has been developed by E.T. Jaynes[4] using the ideas and techniques of information theory, unifies and simplifies statistical mechanics.

5) The view that statistical mechanics is the study of incompletely specified mechanical systems is sometimes criticized on the grounds that it is subjective (i.e. involves the observer), whereas science is supposed to be objective. Without getting into a discussion of whether or not science is actually objective, the author would like to point out that this viewpoint is as objective as any physical theory can be expected to be. According to the operational philosophy of physics, physical theories should depend upon the measuring instruments of the observer. Mechanics deals with the ideal case of perfect and complete measurements. Statistical mechanics makes predictions based on information obtained from imperfect measurements. The predictions naturally depend upon the information, i.e. upon the measuring instruments. The theory does not, however, depend on the subjective views of the observer: two observers, equipped with identical measuring instruments and obtaining identical readings with these instruments, will make precisely the same statistical mechanical predictions about the outcome of future measurements. For example, suppose that a concentration of gaseous U^{235} is diffusing in a box of gaseous U^{238}. Observer A, whose measuring instruments are sensitive only to chemical differences, will say that nothing is happening. Observer B, whose instruments are sensitive to small isotopic mass differences, will say that diffusion is occurring. The description is subjective in the sense that it is different for the two observers, but objective in the sense that the difference depends only on the measuring instruments available. It seems reasonable to expect that the proper statistical mechanical descriptions will be different for observers A and B.

6) The statistical mechanical treatment of systems having only a few degrees of freedom is sometimes criticized on the grounds that predictions then become so highly uncertain that they have no physical significance. But this argument, if accepted, would make it impossible for the physicist to deal with such problems as the hydrogen atom, billiard ball and harmonic oscillator mentioned in the opening paragraph. Furthermore, statistical predictions about small systems do not necessarily have a high uncertainty; whether the system is small or large, the degree of uncertainty in the predictions depends upon the given information.

7) As will be seen in the remainder of this book, the viewpoint that statistical mechanics is the study of incompletely specified mechanical systems leads very naturally to the use, in statistical mechanics, of the mathematical theory of information. It is sometimes objected that this point of view, as developed using information theory, leads to no new physical results. This is not true; a few of the new results obtained using information theory are discussed in Sections 3.6 and 5.4.

A few historical notes might be helpful. J. Willard Gibbs (1839–1903) founded statistical mechanics in its present form (based on the Liouville equation and the distribution function ϱ)[10]. James Clerk Maxwell (1831–1879) and Ludwig Boltzmann (1844–1906) were principal founders of the closely related and sometimes indistinguishable field called *kinetic theory*[11].

In kinetic theory, one tries to derive everything from mechanics, deemphasizing statistical aspects as much as possible; statistical concepts, when introduced, are considered to be merely convenient ways of approximating the exact mechanics. The ultimate aim of kinetic theory is to derive all of the results of statistical mechanics, and in particular the laws governing macroscopic systems, from pure mechanics. In the statistical mechanical approach, on the other hand, one does not try to justify statistical concepts in terms of

mechanics; instead, statistical concepts are considered to be an essential part of the theory, on a par with mechanical concepts.

Kinetic theory, which was well developed by 1865, was originally devised to explain the equilibrium properties of gases in terms of the mechanical behavior of large numbers of molecules. The best known result of this kind is the kinetic theory explanation of the ideal gas laws. Maxwell and Boltzmann extended the theory to include non-equilibrium phenomena in weakly interacting gases, the main results being the famous Boltzmann transport equation and the Boltzmann H-theorem. The emphasis in kinetic theory has always been on "explaining" the macroscopic laws by examining the mechanical behavior of systems having many degrees of freedom. Although it was always necessary to introduce probabilistic assumptions in order to carry out the explanation, it was thought that these assumptions could be removed by a more complete analysis. For example, the justification for the use of many probabilistic assumptions was thought to lie in some form of ergodic theorem, and it was thought that this theorem was likely to hold for large systems. (As noted above, it is now believed that most systems are ergodic, but that this result doesn't depend on the system being large; furthermore, ergodicity is useful only in explaining *equilibrium* phenomena, whereas the Boltzmann transport equation is supposed to describe non-equilibrium phenomena; furthermore, ergodicity does not remove the need for a probabilistic assumption.)

It is typical of kinetic theory that a certain "distribution function" $f(\mathbf{q}, \mathbf{p}, t)$ is used in an ambiguous way: on the one hand, $f(\mathbf{q}, \mathbf{p}, t)\, d^3q\, d^3p$ is interpreted as the *number* of particles having their positions and momenta within the small ranges d^3q, d^3p around the position vector $\mathbf{q}$ and the momentum vector $\mathbf{p}$; on the other hand, $f(\mathbf{q}, \mathbf{p}, t)\, d^3q\, d^3p$ is supposed to be proportional to the *probability* that a particle has its position and momentum inside $d^3q\, d^3p$. Now, the first interpretation says that f is a mechanical quantity, while the second says that it is a probabilistic quantity. As will be shown in Chapter 3, it is inconsistent to interpret the same quantity f in both ways; the distinction between the two interpretations may seem small, but it is crucial. The reader is warned to be on the look-out for this ambiguity when reading the literature, since it arises in nearly every discussion of the Boltzmann transport equation.

Statistical mechanics, in the form presented in most books (i.e. based on the probability distribution function ϱ in phase space) was founded in 1905 with the publication of Gibbs' classic *Elementary Principles in Statistical Mechanics*[10]. Gibbs introduced statistical concepts as a basic ingredient of

the theory with no attempt to explain these concepts mechanically; he thus placed mechanics and statistics on an equal footing. Gibbs considered both equilibrium and non-equilibrium phenomena to be within the scope of his theory, although his treatment of non-equilibrium statistical mechanics is very sketchy.

Gibbs considered his theory to be applicable to any mechanical system, regardless of its size. For example, he states (Ref. 10, p. viii) that "the laws of statistical mechanics apply to conservative systems of any number of degrees of freedom ..." He even explicitly mentions the case of a single particle (Ref. 10, p. 118): "But when $N = 1$, cases may occur in which the canonical distribution is perfectly applicable ...." Thus, Gibbs did not subscribe to the macroscopic viewpoint of statistical mechanics.

Gibbs does not mention the ergodicity concept at any point in his book. In view of the fact that the ergodic hypothesis had been discussed by other writers for over thirty years, it seems safe to say that Gibbs did not consider ergodicity to be relevant to statistical mechanics. Thus, despite the fact that many workers regard ergodicity as a necessary condition for the validity of Gibbs' methods, Gibbs himself did not subscribe to the ergodic viewpoint!

Unfortunately, it is hard to tell from Gibbs' work just how he did justify his own methods. It is clear, however, that he did not subscribe to either the macroscopic viewpoint or the ergodic viewpoint, and that he regarded the introduction of statistical ideas to be a fundamental assumption, not explainable in terms of mechanics.

Since the appearance of Gibbs' work, it has become evident that Gibbs' formalism is more practical and more general than the kinetic theory formalism. In fact, one can treat strongly interacting many-body systems only by using Gibbs' distribution function in the N-body phase space. Thus, except for the specialized case of dilute gases, the Gibbs' formalism has dominated statistical mechanics since 1905. However, the *concepts* (as opposed to the formalism) of kinetic theory have tended to be retained even though Gibbs himself rejected these concepts. Thus, a hybrid theory, using the statistical formalism of Gibbs but the mechanistic concepts of Maxwell and Boltzmann, has developed. As is to be expected, such a situation has led to many paradoxes and to endless debates. The difficulty seems to be that most physicists are, at heart, kinetic theorists, whereas the formalism of Gibbs refuses to yield to a purely kinetic theory interpretation. It is suggested by the author that the appropriate resolution of this situation is that physicists should reorient their intuitions along the lines of a truly *statistical* approach, and that the way to do this is to adopt the view that statistical mechanics is the

study of incompletely specified mechanical systems. This viewpoint seems to be the only one which is both clear and also consistent with the Gibbs formalism. (It was pointed out above that the macroscopic viewpoint does not give a clear interpretation of ϱ, whereas the ergodic viewpoint is inconsistent with the application of Gibbs formalism to non-equilibrium phenomena.)

From 1905 to 1945, the Gibbs formalism (and especially Gibbs' canonical distribution—see Section 3.9) was used to study such equilibrium problems as phase transitions, specific heats, and molecular distribution functions at equilibrium. During this period, most work on non-equilibrium problems was done within the framework of kinetic theory, and was based on the Boltzmann transport equation. This work was valid only for weakly interacting systems (e.g. gases which approximate the ideal gas). During this period, a few workers used the Gibbs formalism (and especially the Liouville equation—see Section 3.3) to study non-equilibrium problems. J. Yvon[12] derived and studied the so-called BBGKY hierarchy; these are the coupled equations for the n-body probability distributions ($n = 1, 2, \ldots, N$) which arise by integrating the Liouville equation over one or more particle coordinates and momenta.

Since 1945, workers have used the Gibbs formalism to carry out more general investigations of non-equilibrium phenomena. J.G. Kirkwood[13] studied the assumptions needed to derive the Boltzmann equation from the Liouville equation. Bogoliubov[14] derived the Boltzmann transport equation by integrating Liouville's equation over the coordinates and momenta of all except one of the particles, making two crucial assumptions about the mathematical behavior of the solution to Liouville's equation, and making certain approximations appropriate to a weakly interacting system. This work set the stage for various extensions of the Boltzmann equation to more general systems. (See Reference 15 for an overview of Bogoliubov's work and related work by others.)

In 1948, Shannon[16] established a new field of mathematics known as **information theory**. Information theory, which is a branch of probability theory, is widely used by communications engineers, and has been applied in such diverse areas as biology and psychology. As its name indicates, information theory is the quantitative study of information: its measurement, and its transmission. If one takes the view that statistical mechanics is the study of incompletely specified mechanical systems, then it becomes natural to try applying the mathematical theory of information to statistical mechanics. E.T. Jaynes[4] seems to have been the first to make a clear, quantitative connection between information theory and statistical mechanics, although several

authors (see, e.g., Refs. 17, 18, and 19) established a qualitative connection prior to the work of Jaynes. A textbook by A. Katz[3] is based on the application of information theory to statistical mechanics. Texts by Landsberg[20] and Tribus[21] develop the relationship of information theory to thermodynamics.

In this book, we will adopt the view that statistical mechanics is the study of incompletely specified systems. In addition, we will use the mathematical theory of information. The relevant portions of probability theory and information theory will be presented in Chapter 2.

REFERENCES

1. R.C. Tolman, *The Principles of Statistical Mechanics* (Oxford University Press, Oxford, 1st edition 1938).
2. F.C. Andrews, *Equilibrium Statistical Mechanics* (John Wiley and Sons, Inc., New York, 1963).
3. A. Katz, *Principles of Statistical Mechanics: The Information Theory Approach* (W.H. Freeman and Company, San Francisco, 1967).
4. E.T. Jaynes, *Phys. Rev.* **106**, 620 (1957); *Phys. Rev.* **108**, 171 (1957); in *Statistical Physics* (1962 Brandeis Lectures), ed. by K.W. Ford (W.A. Benjamin, Inc., New York, 1963); *Am. J. Phys.* **33**, 391 (1965); in *Delaware Seminar in the Foundations of Physics*, Vol. 1, ed. by M. Bunge (Springer-Verlag, New York, 1967).
5. H. Grad, in *Delaware Seminar in the Foundations of Physics*, Vol. 1, ed. by M. Bunge (Springer-Verlag, New York, 1967).
6. C. Truesdale, in *Ergodic Theories*, Proc. of the International School of Physics "Enrico Fermi", course 14 (Academic Press, New York, 1961).
7. Ja. G. Sinai, in *Statistical Mechanics*, Proc. of the I.U.P.A.P. Meeting, Copenhagen, 1966, ed. by T.A. Bak (W.A. Benjamin, Inc., New York, 1967).
8. I.E. Farquhar, *Ergodic Theory in Statistical Mechanics* (Interscience Publishers, New York, 1964).
9. I.E. Farquhar, in *Statistical Mechanics*, Proc. of the I.U.P.A.P. Meeting, Copenhagen, 1966, ed. by T.A. Bak (W.A. Benjamin, Inc., New York, 1967).
10. J.W. Gibbs, *Elementary Principles in Statistical Mechanics* (Dover Publications, Inc., New York, 1960; this edition is an unabridged republication of the work first published by the Yale University Press in 1905).
11. S.G. Brush, *Kinetic Theory*, Vols. 1 and 2 (Pergamon Press, Oxford, 1966).
12. J. Yvon, *Actualities Scientifiques et Industriel* (Herman et Cie, Paris, 1935).
13. J.G. Kirkwood, *J. Chem. Phys.* **14**, 180 (1946); **15**, 72 (1947).
14. N.N. Bogoliubov, in *Studies in Statistical Mechanics*, Vol. 1, ed. by J. de Boer and G.E. Uhlenbeck (North-Holland Publishing Company, Amsterdam, 1962). (This is an English translation of the original 1946 article).
15. T. Wu, *Kinetic Equations of Gases and Plasmas* (Addison-Wesley Publishing Company, Reading, Massachusetts, 1966).
16. C.E. Shannon, *Bell System Tech. J.* **27**, 379 and 623 (1948). Reprinted in C.E. Shannon, and W. Weaver, *The Mathematical Theory of Communication* (Univ. of Illinois Press, Urbana, 1949).

17. L. Brillouin, *J. of Appl. Phys.* **22**, 334 and 338 (1951); *Science and Information Theory* (Academic Press, New York, 1956).
18. R. C. Raymond, *Am. J. Phys.* **19**, 109 (1951).
19. J. Rothstein, *Science* **114**, 171 (1951); *J. Appl. Phys.* **23**, 1281 (1952); *Phys. Rev.* **85**, 135 (1952).
20. P. T. Landsberg, *Thermodynamics* (Interscience Publishers, New York, 1961).
21. M. Tribus, *Thermostatics and Thermodynamics* (D. Van Nostrand Company, Princeton, 1961).

CHAPTER 2

Probability

2.1 THE MATHEMATICS OF PROBABILITY

THE DISTINCTION between mathematics on the one hand, and its applications or interpretations on the other, seems clear in most fields. It is not difficult, for example, to distinguish the formal theory of second order differential equations from its applications in mechanics, circuit analysis, etc. Unfortunately, this distinction is not so clear in the case of probability theory. This is partly due to the wide use of the language of probability theory in everyday language ("random", "chance", "event", "probable", "certain", etc.), and is partly due to the close historical connection between the mathematical theory and applications to such areas as games of chance, population studies, communications engineering, and statistical mechanics. In order to keep the distinction between the formalism and its interpretations clear, we will present the mathematics of probability separately from the interpretations of probability. This section is about the easy part: the mathematics. We will discuss the interpretations in the next section.

For concreteness, we will illustrate the formalism by five simple examples. We present these examples not for the purpose of clarifying the *meaning* of probability (this will be done, hopefully, in the next section), but merely to aid the reader in grasping the mathematics. Since several of these examples run throughout this section, the examples will be listed here for reference:

Example 1: One throw of a single die, with "even" or "odd" being the event of interest.
Example 2: One throw of a single die, with the number of spots being the event of interest.
Example 3: Choosing "at random" a number between 0 and 1.
Example 4: One throw of two dice.
Example 5: Ten throws of a single die.

Set Notation. We assume that the reader has some familiarity with set operations. The notation will be as follows:

$\{\zeta_1, \zeta_2 \ldots\}$ means the set whose elements, or members, are $\zeta_1, \zeta_2, \ldots$ Sets need not be finite, and they need not be countable.

$\mathscr{S}$ means the universal set, i.e. the set containing all those elements ζ under consideration in a given discussion.

$\mathscr{O}$ means the null set, containing no elements.

$\mathscr{E}$ or $\mathscr{E}_1$ means a non-null subset of $\mathscr{S}$.

$\zeta \in \mathscr{E}$ means that the element ζ is a member of, or belongs to, the set $\mathscr{E}$.

$\zeta \notin \mathscr{E}$ means that ζ is not a member of $\mathscr{E}$.

$\mathscr{E}_1 \subset \mathscr{E}_2$, or $\mathscr{E}_2 \supset \mathscr{E}_1$, means that $\mathscr{E}_1$ is a subset of, or is contained in, $\mathscr{E}_2$. That is, every $\zeta \in \mathscr{E}_1$ satisfies $\zeta \in \mathscr{E}_2$.

$\bar{\mathscr{E}}$ means the compliment of $\mathscr{E}$, i.e. the set consisting of all the elements of $\mathscr{S}$ satisfying $\zeta \notin \mathscr{E}$.

$\mathscr{E}_1 \cup \mathscr{E}_2$ means the union of $\mathscr{E}_1$ and $\mathscr{E}_2$, i.e. the set of all ζ such that either $\zeta \in \mathscr{E}_1$ or $\zeta \in \mathscr{E}_2$ or both.

$\mathscr{E}_1 \cap \mathscr{E}_2$ means the intersection of $\mathscr{E}_1$ and $\mathscr{E}_2$, i.e. the set of all ζ such that $\zeta \in \mathscr{E}_1$ and $\zeta \in \mathscr{E}_2$.

Finally, note the distinction between the element ζ_1 and the set $\{\zeta_1\}$. We write $\zeta_1 \in \mathscr{S}$, but $\{\zeta_1\} \subset \mathscr{S}$. A set, such as $\{\zeta_1\}$, containing a single element is sometimes called an "elementary set".

The Structure of Probability Theory. The theory is based on the abstract notion of a **random experiment**. A particular random experiment is defined by specifying a **sample space** $\mathscr{S}$, **field of events** $\mathbf{F}$, and a **set of probabilities** P defined on $\mathbf{F}$. Thus, we speak of the experiment $(\mathscr{S}, \mathbf{F}, P)$. The postulated properties of $\mathscr{S}$, $\mathbf{F}$ and P will now be given. Note that an experiment $(\mathscr{S}, \mathbf{F}, P)$ is a purely abstract, mathematical concept, and does not necessarily have any relationship to the word "experiment" as used in science or in everyday language.

Sample Spaces. A **sample space** $\mathscr{S}$ is a set of elements $\zeta_1, \zeta_2, \ldots$ This set need not be finite or countable. The elements $\zeta \in \mathscr{S}$ are called **outcomes**.

Example 1. Suppose that the experiment is the throwing of a single die. A natural sample space is the set $\mathscr{S} = \{f_1, f_2, f_3, f_4, f_5, f_6\}$, where the outcome f_i means "the throw resulted in i spots".

Fields of Events. A **field of events** means a collection $\mathbf{F}$ of subsets $\mathscr{E}_1, \mathscr{E}_2, \ldots$ of $\mathscr{S}$, such that

1. If $\mathscr{E} \in \mathbf{F}$, then $\bar{\mathscr{E}} \in \mathbf{F}$.
2. If $\mathscr{E}_1 \in \mathbf{F}$ and $\mathscr{E}_2 \in \mathbf{F}$, then $\mathscr{E}_1 \cup \mathscr{E}_2 \in \mathbf{F}$.

That is, $\mathbf{F}$ means a collection of subsets (i.e. a "set of sets") of $\mathscr{S}$, such that this collection has the properties 1 and 2. The members of $\mathbf{F}$ are called

events $\mathscr{E}$. Thus, any event $\mathscr{E} \in \mathbf{F}$ is a subset of $\mathscr{S}$. However, it is not necessarily true that every subset of $\mathscr{S}$ is an event, i.e. is a member of $\mathbf{F}$. It can be shown[1] that the field postulates 1 and 2 imply:

3. If $\mathscr{E}_1 \in \mathbf{F}$ and $\mathscr{E}_2 \in \mathbf{F}$, then $\mathscr{E}_1 \cap \mathscr{E}_2 \in \mathbf{F}$.
4. $\mathscr{S} \in \mathbf{F}$, i.e. $\mathscr{S}$ is itself an event, called the **certain event**.
5. $\mathscr{O} \in \mathbf{F}$, i.e. $\mathscr{O}$ is an event, called the **impossible event.**

Properties 2 and 3 imply that every union or intersection of a finite number of events is a member of $\mathbf{F}$. However, if $\mathbf{F}$ contains an infinite number of events, then 2 and 3 do not insure that every infinite union or intersection of events is a member of $\mathbf{F}$. In this case we postulate that $\mathbf{F}$ is a Borel field, i.e. we add to postulates 1 and 2 the postulate that $\mathbf{F}$ contains every infinite union and intersection of its members.

Example 1 (continued). Suppose that we are not interested in the precise number of spots, but are interested only in whether the throw is even or odd. A natural choice for $\mathbf{F}$ is then $\mathbf{F} = \{\mathscr{O}\, \mathscr{E}'\, \mathscr{E}''\, \mathscr{S}\}$, where $\mathscr{E}' = \{f_1 f_3 f_5\}$ is the "odd" event, and $\mathscr{E}'' = \{f_2 f_4 f_6\}$ is the "even" event. This collection $\mathbf{F}$ of four events satisfies the field postulates 1 and 2. Note that $\bar{\mathscr{E}}' = \mathscr{E}''$, $\mathscr{E}' \cup \mathscr{E}'' = \mathscr{S}$, $\mathscr{E}' \cap \mathscr{E}'' = \mathscr{O}$. In applications, any event is said to **occur** on a certain trial if the outcome ζ of the trial satisfies $\zeta \in \mathscr{E}$. If the die is thrown and the outcome is f_3, then $\mathscr{E}'$ and $\mathscr{S}$ occurred whereas $\mathscr{E}''$ and $\mathscr{O}$ did not occur. Note that $\mathscr{S}$ occurs on every trial, while $\mathscr{O}$ never occurs.

Probabilities. A **set of probabilities** P means a numerical function of the events $\mathscr{E}$, defined for every $\mathscr{E} \in \mathbf{F}$, and satisfying the following conditions:

$$P(\mathscr{E}) \geqslant 0 \quad \text{for every} \quad \mathscr{E} \in \mathbf{F}, \tag{2.1.1}$$

$$P(\mathscr{S}) = 1, \tag{2.1.2}$$

$$P(\mathscr{E}_1 \cup \mathscr{E}_2) = P(\mathscr{E}_1) + P(\mathscr{E}_2) \text{ whenever } \mathscr{E}_1 \cap \mathscr{E}_2 = 0. \tag{2.1.3}$$

Note that $P(\mathscr{E})$ is a function of events, i.e. for every event $\mathscr{E} \in \mathbf{F}$, a number $P(\mathscr{E})$ is assigned. A slight extension of postulate (3) is required in case $\mathbf{F}$ contains an infinite number of events;[1] this extension will not be given here. Postulates (1), (2) and (3) imply

$$0 \leqslant P(\mathscr{E}) \leqslant 1, \tag{2.1.4}$$

$$P(\mathscr{O}) = 0, \tag{2.1.5}$$

$$P(\mathscr{E}_1 \cup \mathscr{E}_2) = P(\mathscr{E}_1) + P(\mathscr{E}_2) - P(\mathscr{E}_1 \cap \mathscr{E}_2), \tag{2.1.6}$$

$$P(\mathscr{E}) + P(\bar{\mathscr{E}}) = 1. \tag{2.1.7}$$

For the proofs, see Ref. 1.

Example 2. Let the experiment again be the throwing of a single die, but suppose that we are now interested in which face comes up. The natural prescription of the experiment $(\mathscr{S}, \mathbf{F}, P)$ is: $\mathscr{S} = \{f_1 f_2 \cdots f_6\}$; $\mathbf{F}$ is the set of all possible subsets of $\mathscr{S}$; $P(\mathscr{E}) = n/6$ where n is the number of outcomes in $\mathscr{E}$. For instance, the probability that the number of spots will be less than or equal to 2 is $P(\{f_1 f_2\}) = 2/6 = 1/3$. Nothing in the mathematical formalism requires us to use this particular probability assignment P; this assignment seems, however, to be the most "reasonable" to most people. The reader should convince himself that this assignment of $(\mathscr{S}, \mathbf{F}, P)$ satisfies the field postulates 1 and 2, as well as the probability postulates (1), (2), and (3). As an illustration of (6), note that (6) implies

$$\begin{aligned} P(\{f_1 f_2\} \cup \{f_2 f_3\}) &= P(\{f_1 f_2\}) + P(\{f_2 f_3\}) - P(\{f_1 f_2\} \cap \{f_2 f_3\}) \\ &= P(\{f_1 f_2\}) + P(\{f_2 f_3\}) - P(\{f_2\}) \\ &= 2/6 + 2/6 - 1/6 = 1/2. \end{aligned}$$

The same result follows more directly from the fact that

$$\{f_1 f_2\} \cup \{f_2 f_3\} = \{f_1 f_2 f_3\}.$$

Conditional Probability. The **conditional probability** of $\mathscr{E}_2$ assuming $\mathscr{E}_1$, also called the probability of $\mathscr{E}_2$ given $\mathscr{E}_1$, is defined by

$$P(\mathscr{E}_2 | \mathscr{E}_1) = \frac{P(\mathscr{E}_2 \cap \mathscr{E}_1)}{P(\mathscr{E}_1)} \tag{2.1.8}$$

provided $P(\mathscr{E}_1) \neq 0$, and is undefined if $P(\mathscr{E}_1) = 0$. Thus

$$P(\mathscr{E}_2 \cap \mathscr{E}_1) = P(\mathscr{E}_2 | \mathscr{E}_1) P(\mathscr{E}_1), \tag{2.1.9}$$

which says that the probability that both $\mathscr{E}_1$ and $\mathscr{E}_2$ will occur equals the probability of $\mathscr{E}_2$ given $\mathscr{E}_1$, times the probability that $\mathscr{E}_1$ will occur. If $\mathscr{E}_1 \subset \mathscr{E}_2$ then (8) implies $P_2(\mathscr{E}_2 | \mathscr{E}_1) = 1$, which says that the probability of any event $\mathscr{E}_2$ implied by $\mathscr{E}_1$, given that $\mathscr{E}_1$ occurred, is 1. If on the other hand $\mathscr{E}_1 \cap \mathscr{E}_2 = 0$, then $P(\mathscr{E}_2 | \mathscr{E}_1) = 0$, which says that the probability of any event $\mathscr{E}_2$ excluded by $\mathscr{E}_1$, given that $\mathscr{E}_1$ occurred, is 0.

Example 2 (continued). From (8),

$$P(\{f_1 f_2\} \mid \text{odd}) = \frac{P(\{f_1 f_2\} \cap \text{odd})}{P(\text{odd})} = \frac{P(\{f_1\})}{P(\text{odd})} = \frac{1/6}{1/2} = \frac{1}{3},$$

$$P(\text{odd} \mid \{f_1\}) = 1,$$

$$P(\text{odd} \mid \{f_2\}) = 0.$$

Expressions such as $P(\mathscr{E} \mid \text{odd} \cap \{f_2\})$ are undefined, since $\text{odd} \cap \{f_2\}$ is the impossible event.

Statistical Independence. Two events $\mathscr{E}_1$ and $\mathscr{E}_2$ are **statistically independent** if

$$P(\mathscr{E}_1 \cap \mathscr{E}_2) = P(\mathscr{E}_1) P(\mathscr{E}_2), \tag{2.1.10}$$

i.e. if the probability of both events occurring equals the product of the probability that $\mathscr{E}_1$ will occur and the probability that $\mathscr{E}_2$ will occur. From (8), statistical independence of $\mathscr{E}_1$ and $\mathscr{E}_2$ implies

$$P(\mathscr{E}_1 | \mathscr{E}_2) = P(\mathscr{E}_1), \quad \text{and} \quad P(\mathscr{E}_2 | \mathscr{E}_1) = P(\mathscr{E}_2). \tag{2.1.11}$$

The first of these relations says that the probability of $\mathscr{E}_1$ given that $\mathscr{E}_2$ occurred equals the probability of $\mathscr{E}_1$ alone, i.e. the probability of $\mathscr{E}_1$ is the same whether or not it is given that $\mathscr{E}_2$ occurred. If $\mathscr{E}_1$ and $\mathscr{E}_2$ are not statistically independent, they are said to be **correlated**.

The word *independence* has many distinct meanings in science and mathematics: linear independence, independent variables, dynamical independence, statistical independence, etc. In statistical mechanics, statistical independence is sometimes confused with other types of independence. The reader is advised to use this word with care.

Example 2 (continued). The events $\{f_1\}$ and "odd" are correlated, since $P(\{f_1\} | \text{odd}) = 1/3$ while $P(\{f_1\}) = 1/6$. In fact, in this experiment two events are statistically independent only if one of them is $\mathscr{S}$ or $\mathscr{O}$.

Random Variables. A **random variable** (r.v.) is any numerical function of the outcomes ζ of a random experiment.* Random variables will be denoted by capital letters $X, Y, \ldots$ Thus anr.v. X is a relationship between the outcomes ζ of a random experiment $(\mathscr{S}, \mathbf{F}, P)$ and the corresponding numbers $X(\zeta)$. Note that probabilities are numerical functions of the events $\mathscr{E}$, defined for every $\mathscr{E} \in \mathbf{F}$, while anr.v. is a numerical function of the outcomes ζ, defined for every $\zeta \in \mathscr{S}$. The notation $\{X(\zeta) = x\}$, or simply $\{X = x\}$, will mean the *event* consisting of all the elements ζ satisfying the condition that $X(\zeta)$ equals the number x. Note that $X\{\zeta\}$ is a *number*, but that $\{X(\zeta) = x\}$ is an *event*, i.e. a set of outcomes. An extension of this notation is illustrated in the following example.

Example 2 (continued). Define the r.v. X and Y by

$$X(f_i) = i, \tag{2.1.12}$$

$$Y(f_i) = \begin{cases} 0 & (i = 1, 2) \\ 1 & (i = 3, 4, 5, 6). \end{cases} \tag{2.1.13}$$

* More precisely, a r.v. is any numerical function $X(\zeta)$ which is "measurable with respect to the field $\mathbf{F}$", i.e. which satisfies the condition that $\{X \leqslant x\}$ be an event for every choice of the number x. In practice, all "reasonable" functions $X(\zeta)$ satisfy this requirement.

Then the events $[X = 3]$, $[X \leqslant 2]$, $[X \text{ odd}]$, $[Y = 1]$ are given by

$$[X = 3] = [f_3],$$
$$[X \leqslant 2] = [f_1 f_2] = [Y = 0],$$
$$[X \text{ odd}] = [f_1 f_3 f_5],$$
$$[Y = 1] = [f_3 f_4 f_5 f_6].$$

A **discrete random variable** is one whose range (i.e. the set of numbers $X(\zeta)$ corresponding to all possible outcomes ζ) is entirely discrete. The r.v. (12) and (13) are discrete. A **continuous random variable** is one whose range is entirely continuous. In this book we will consider only discrete or continuous r.v.

Example 3. Suppose we pick a number "at random" from the interval $0 \leqslant \zeta \leqslant 1$ of the real axis. For this experiment, the natural choice of $(\mathscr{S}, \mathbf{F}, P)$ is: $\mathscr{S}$ is the set of real numbers ζ between 0 and 1; $\mathbf{F}$ consists of all Lebesgue measurable* subsets of $\mathscr{S}$; $P(\mathscr{E})$ is the Lebesgue measure* of the set $\mathscr{E}$. Define the r.v. X, Y, U, V by

$$X(\zeta) = \zeta, \tag{2.1.14}$$
$$Y(\zeta) = \zeta^2, \tag{2.1.15}$$
$$U(\zeta) = \begin{cases} 0 & (0 \leqslant \zeta \leqslant 0.5) \\ 1 & (0.5 < \zeta \leqslant 1) \end{cases} \tag{2.1.16}$$
$$V(\zeta) = \begin{cases} n & \left(\dfrac{1}{n+1} < \zeta \leqslant \dfrac{1}{n};\ \ n = 1, 2, \ldots\right) \\ 0 & (\zeta = 0). \end{cases} \tag{2.1.17}$$

Then X and Y are continuous r.v., while U and V are discrete.

Functions of Random Variables. Consider an r.v. X and a numerical function $g(x)$, such that $g(x)$ is defined for every number x in the range of X. Corresponding to the outcome ζ, X takes on the numerical value $X(\zeta)$. We

* Readers unfamiliar with measure theory need not be alarmed by the terms "Lebesgue measurable" and "Lebesgue measure". In practice, the term "Lebesgue measurable" may be replaced by "reasonable", and the term "Lebesgue measure" may be replaced by "volume" or "length"; readers unfamiliar with measure theory are urged to make this replacement wherever these terms occur. It is worth noting that every open or closed interval on the real axis is measurable, and has a measure equal to the length of the interval; every discrete set of points on the real axis is measurable and has measure zero. The reason for the requirement that $\mathbf{F}$ consist only of the Lebesgue measurable sets (rather than *all* sets) is that this insures that it is possible to define on $\mathbf{F}$ a set of probabilities satisfying postulates (1), (2), and (3).

then define the function $g(X)$ of the r.v. X by the following statement: corresponding to the outcome ζ, the numerical value of $g(X)$ is $g(X(\zeta))$. Note that $g(X)$ assigns a number to every outcome ζ, and is thus a random variable.

Example 2 (continued). Defining X by (12), the new r.v. $U = X^2$ takes on the values $U(f_1) = 1$, $U(f_2) = 4$, $U(f_3) = 9$, etc.

Example 3 (continued). The r.v. defined by (14) and (15) are related by $Y = X^2$. The r.v. defined by (16) and (17) are related by $U = g(V)$ where $g(V) = 0$ for $V = 0, 2, 3, 4, \ldots$, and $g(1) = 1$.

Probability Distributions. The probability distribution of an r.v. X is, roughly, a statement giving the probabilities associated with the various values of X. We will consider the cases of discrete and continuous r.v. separately.

(i) Discrete r.v. The **probability distribution** of a discrete r.v. X is the set of probabilities of the events $\{X = x_i\}$, for the various values x_i in the range of X. We will denote the probability distribution in any of several ways:

$$P(\{X = x_i\}) = p(x_i) = p_i. \tag{2.1.8}$$

Example 2 (continued). The probability distribution of the r.v. Y defined by (13) is

$$p(0) = P(\{Y = 0\}) = P(\{f_1 f_2\}) = 1/3,$$
$$p(1) = P(\{Y = 1\}) = P(\{f_3 f_4 f_5 f_6\}) = 2/3.$$

(ii) Continuous r.v. The **probability density** (or **probability distribution**) of a continuous r.v. X means the function

$$\varrho(x) \equiv \lim_{\Delta x \to 0} \frac{P(\{x < X \leqslant x + \Delta x\})}{\Delta x}. \tag{2.1.19}$$

Thus, for small Δx, $\varrho(x)\,\Delta x$ is approximately the probability of the event $\{x < X \leqslant x + \Delta x\}$. By (19) and (1), $\varrho(x)$ is non-negative.

Example 3 (continued). Defining the r.v. Y by (15), we have

$$P(\{y < Y \leqslant y + \Delta y\}) = P(\{y < \zeta^2 \leqslant y + \Delta y\})$$
$$= P(\{\sqrt{y} < \zeta \leqslant \sqrt{y + \Delta y}\}) = \sqrt{y + \Delta y} - \sqrt{y}.$$

Using (19),

$$\varrho(y) = \lim_{\Delta y \to 0} \frac{\sqrt{y + \Delta y} - \sqrt{y}}{\Delta y} = \frac{1}{2\sqrt{y}}. \tag{2.1.20}$$

Warning: The words "distribution" and "density" are used in many ways: mass distribution, particle distribution, velocity distribution, and in addition,

probability distribution. Unfortunately, these radically different concepts have certain similarities (e.g. the "velocity distribution" is superficially like the "probability distribution for velocities"), and there is a widespread tendency to get them confused. Since probabilistic concepts are at the heart of statistical mechanics, it is essential, if one is to have any hope of grasping the subject, to maintain the distinction between these different ideas.

Normalization.

(i) Discrete r.v. The events $\{X = x_i\}$ are disjoint, i.e.

$$\{X = x_i\} \cap \{X = x_j\} = 0 \quad (i \neq j).$$

Furthermore, the event $\{X$ takes on one of the values x_i in its range$\}$ is the certain event $\mathscr{S}$. Thus, by postulates (2) and (3),

$$\begin{aligned} 1 = P(\mathscr{S}) &= P(\{X \text{ takes on one of the values } x_i\}) \\ &= P(\{X = x_1\} \cup \{X = x_2\} \cup \cdots) \\ &= P(\{X = x_1\}) + P(\{X = x_2\}) + \cdots, \end{aligned}$$

or

$$\Sigma\, p(x_i) = 1, \tag{2.1.21}$$

where the sum is taken over all values x_i in the range of X. Equation (21) is the **normalization** condition for discrete r.v. Using similar reasoning, we see that the probability that X will take on one of the values x_i in an arbitrary set M of possible values is

$$P(\{X(\zeta) \in M\}) = \sum_{x_i \in M} p(x_i). \tag{2.1.22}$$

Example 3 (continued). The probability distribution for the r.v. V defined by (17) is

$$\left.\begin{aligned} p(n) = P(\{V = n\}) &= \frac{1}{n} - \frac{1}{n+1} \quad (n = 1, 2, \ldots) \\ p(0) = P(\{V = 0\}) &= 0. \end{aligned}\right\} \tag{2.1.23}$$

Using (23), we see explicitly that

$$\sum_{n=0}^{\infty} p(n) = 1.$$

As an example of (22),

$$P(\{V \leqslant 2\}) = p(0) + p(1) + p(2) = 2/3.$$

(ii) Continuous r.v. By an argument similar to that used in obtaining (21), we can show that the probability distribution of any continuous r.v. X must satisfy the normalization condition

$$\int \varrho(x)\, dx = 1, \tag{2.1.24}$$

where the integral is taken over the entire range of X. Analogously to (22),

$$P(\{x_1 \leqslant X \leqslant x_2\}) = \int_{x_1}^{x_2} \varrho(x)\, dx, \tag{2.1.25}$$

from which $P(\{X = x\}) = 0$ for *any* number x in the range of the continuous r.v. X.

Example 3 (continued). We can explicitly check the normalization of the probability distribution (20):

$$\int_0^1 \frac{dy}{2\sqrt{y}} = 1.$$

Using (25) and (20), with Y given by (15),

$$P(\{0 \leqslant Y \leqslant 0.5\}) = \int_0^{0.5} \varrho(y)\, dy = 0.71,$$

$$P(\{0.5 < Y \leqslant 1\}) = \int_{0.5}^1 \varrho(y)\, dy = 0.29.$$

Probability Distributions for N Random Variables. Consider two r.v. X and Y, defined on the *same* experiment $(\mathscr{S}, \mathbf{F}, P)$. (Henceforth, it will be assumed that all r.v. under consideration are defined on the same experiment $(\mathscr{S}, \mathbf{F}, P)$.)

(i) Discrete r.v. Consider the events $\{X = x_i\} \cap \{Y = y_j\}$, i.e. "$X$ takes the value x_i and Y takes the value y_j". The **joint probability distribution** of X and Y means the set of probabilities of these joint events. Any of the following notations may be used:

$$P(\{X = x_i\} \cap \{Y = y_j\}) = p(x_i, y_j) = p_{ij}. \tag{2.1.26}$$

The extension to the case of N r.v. $X, Y, \ldots, Z$ is straightforward; in this case, the probability distribution is written $p(x_i, y_j, \ldots, z_k)$ or $p_{ij\ldots k}$.

Example 2 (continued). We can find the joint distribution of the r.v. (12) and (13) as follows: By (9),

$$P(\{X = x\} \cap \{Y = y\}) = P(\{X = x\} \mid \{Y = y\})\, P(\{Y = y\}).$$

Thus

$$p(x, y) = \begin{cases} (1/2)(1/3) = 1/6 & (x = 1, 2; \quad y = 0) \\ (0)(1/3) = 0 & (x = 3, 4, 5, 6; \quad y = 0) \\ (0)(2/3) = 0 & (x = 1, 2; \quad y = 1) \\ (1/4)(2/3) = 1/6 & (x = 3, 4, 5, 6; \quad y = 1), \end{cases}$$

(ii) Continuous r.v. Consider the events $\{x < X \leqslant x + \Delta x\} \cap \{y < Y \leqslant y + \Delta y\}$, and define the **joint probability density** by

$$\varrho(x, y) \equiv \lim_{\Delta x \to 0} \lim_{\Delta x \to 0} \frac{P(\{x < X \leqslant x + \Delta x\} \cap \{y < Y \leqslant y + \Delta y\})}{\Delta x \, \Delta y}. \tag{2.1.27}$$

Thus, $\varrho(x, y)\, \Delta x \, \Delta y$ is (approximately) the probability that X is in the range $x < X \leqslant x + \Delta x$, *and* Y is in the range $y < X \leqslant y + \Delta y$. The extension to N r.v. is straightforward.

Example 3 (continued). Defining X and Y by (14) and (15), we see that the joint distribution satisfies

$$\varrho(x, y) = 0 \quad \text{for} \quad y \neq x^2. \tag{2.1.28}$$

Thus $\varrho(x, y) \propto \delta(y - x^2)$, where $\delta(x)$ is the generalized function known as the "Dirac delta function".[2] Choosing the proportionality constant in such a way that $\varrho(x, y)$ is normalized on $(0 \leqslant x \leqslant 1,\ 0 \leqslant y \leqslant 1)$, we obtain

$$\varrho(x, y) = \delta(y - x^2). \tag{2.1.29}$$

The form $\varrho(x, y) \propto \delta(y - f(x))$ is typical of "functionaly related" r.v., i.e. continuous r.v. related by $Y = f(X)$.

Reduced Probability Distributions. Given two r.v. X and Y with joint probability distribution $p(x_i, y_j)$ or $\varrho(x, y)$, it is common to call the probability distribution for X alone or for Y alone a **reduced distribution** obtained from the joint distribution. As will now be shown, we can express any reduced distribution in terms of the joint distribution.

(i) Discrete r.v. Using postulate (3),

$$\begin{aligned} P(\{X = x_i\}) &= P(\{X = x_i\} \cap \{Y \text{ takes on one of its values}\}) \\ &= \sum_{y_j} p_{XY}(x_i, y_j). \end{aligned}$$

Thus, the probability distribution for X alone is

$$p_X(x_i) = \sum_{y_j} p_{XY}(x_i, y_j). \tag{2.1.30}$$

(Concerning the notation: whenever there is any possibility of confusion, we will distinguish probability distributions for different r.v. by appropriate subscripts. For example, we denote the distribution for X by $\varrho_X(\ )$.) According to (30), we "reduce" p_{XY} (i.e. sum over all values of Y) in order to obtain p_X. The extension to N discrete r.v. is straightforward.

(ii) Continuous r.v. By the reasoning used in obtaining (30),

$$\varrho_X(x) = \int \varrho_{XY}(x, y)\, dy, \tag{2.1.31}$$

where the integral is taken over the range of Y.

Example 3 (continued). The joint distribution (29) for the r.v. (14) and (15) leads to

$$\begin{aligned} \varrho_Y(y) &= \int_0^1 \delta(y - x^2)\, dx = \int_0^1 \delta[(x - \sqrt{y})(x + \sqrt{y})]\, dx \\ &= \int_0^1 (x + \sqrt{y})^{-1}\, \delta(x - \sqrt{y})\, dx = (\sqrt{y} + \sqrt{y})^{-1} \\ &= 1/2\sqrt{y}, \end{aligned}$$

in agreement with (20).

Statistical Independence of Random Variables.

(i) Discrete r.v. Two discrete r.v. X and Y are **statistically independent** if the pair of events $\{X = x_i\}$, $\{Y = y_j\}$ is statistically independent for every x_i and y_j. Thus, using (10), the condition for statistical independence of X and Y is

$$P(\{X = x_i\} \cap \{Y = y_j\}) = P(\{X = x_i\})\, P(\{X = y_j\}), \tag{2.1.32}$$

or

$$p_{XY}(x_i, y_j) = p_X(x_i)\, p_Y(y_j), \tag{2.1.33}$$

for all (x_i, y_j). By (11), statistical independence of X and Y implies

$$P(\{X = x_i\} \mid \{Y = y_j\}) = P(\{X = x_i\}) \tag{2.1.34}$$

for all (x_i, y_j). This says that the probability distribution for X is the same whether or not it is given that $\{Y = y_j\}$ occurred. We say that two r.v. are **correlated** if they are not statistically independent. Note that X and Y are statistically independent only if (32) holds for *all* x_i and y_j; if (32) holds for some (x_i, y_j) but not for others, then X and Y are correlated.

Example 2 (continued). The r.v. defined by (12) and (13) are correlated, since

$$P(\{X = 5\} \mid \{Y = 0\}) = 0 \quad \text{and} \quad P(\{X = 5\}) = 1/6.$$

Example 4. Two dice are thrown. The experiment $(\mathscr{S}, \mathbf{F}, P)$ is traditionally defined as follows: The outcomes, denoted f_{ij}, are "the first die shows i spots and the second die shows j spots". The sample space $\mathscr{S}$ is the set $\{f_{11} f_{12} \cdots f_{21} f_{22} \cdots f_{66}\}$. The field $\mathbf{F}$ consists of all subsets of $\mathscr{S}$. The probabilities of the elementary events $\{f_{ij}\}$ are 1/36, and the probabilities of all other events then follow from postulate (3); for example,

$$\begin{aligned} P(\{f_{11} f_{12}\}) &= P(\{f_{11}\} \cup \{f_{12}\}) \\ &= P(\{f_{11}\}) + P(\{f_{12}\}) = 1/36 + 1/36 = 1/18. \end{aligned}$$

For this experiment, it is convenient to define the r.v.

$$X(f_{ij}) = i, \quad Y(f_{ij}) = j, \tag{2.1.35}$$

i.e. X is the number of spots showing on the first die, and Y is the number showing on the second. The joint distribution of X and Y is $p(x_i, y_j) = 1/36$ $(i, j = 1, 2, \ldots, 6)$. Since $P(\{X = x_i\} \mid \{Y = y_j\}) = 1/6 = P(\{X = x_i\})$, the r.v. defined by (35) are statistically independent. Thus, the two dice are uncorrelated. If the two dice were connected in some way, for instance by gluing them together, the probability assignment given above would not seem reasonable to most people. A reasonable probability assignment for the connected dice would obviously lead to correlations between X and Y.

(ii) Continuous r.v. Two continuous r.v. are **statistically independent** if the pair of events $\{x < X \leqslant x + \Delta x\}$, $\{y < Y \leqslant y + \Delta y\}$ is statistically independent for every choice of x, y, Δx, and Δy (Δx and Δy need not be small). Thus, the condition for statistical independence is

$$P(\{x < X \leqslant x + \Delta x\} \cap \{y < Y \leqslant y + \Delta y\}) = P(\{x < X \leqslant x + \Delta x\})\, P(\{y < Y \leqslant y + \Delta y\}). \tag{2.1.36}$$

Using (19) and (27), equation (36) implies

$$\varrho_{XY}(x, y) = \varrho_X(x)\, \varrho_Y(y). \tag{2.1.37}$$

Equation (37) in turn implies (36), so that (37) is the necessary and sufficient condition for statistical independence of X and Y. Statistical independence of X and Y implies

$$P(\{x < X \leqslant x + \Delta x\} \mid \{y < Y \leqslant y + \Delta y\}) = P(\{x < X \leqslant x + \Delta x\}), \tag{2.1.38}$$

which says, roughly, that the probability distribution for X is the same regardless of what is given about Y.

Example 3 (continued). The r.v. (14) and (15) are correlated, since the value of X *determines* Y. Cases of this type, in which Y is functionally related to X, are the extreme opposite of statistical independence.

Expectation Values and Moments. We define the **expectation value** of a function $g(X)$ of the r.v. X by

$$\langle g(X) \rangle = \sum_i g(x_i)\, p(x_i) \tag{2.1.39}$$

for discrete X, and by

$$\langle g(X) \rangle = \int g(x)\, \varrho(x)\, dx \tag{2.1.40}$$

for continuous X. The operation $\langle\ \rangle$ is a *linear* operation, i.e. for arbitrary numbers a_i and functions g_i,

$$\langle a_1 g_1(X) + a_2 g_2(X) \rangle = a_1 \langle g_1(X) \rangle + a_2 \langle g_2(X) \rangle. \tag{2.1.41}$$

The nth **moment** m_n of X means the expected value of X^n:

$$m_n = n\text{th moment of } X = \langle X^n \rangle. \tag{2.1.42}$$

The nth **central moment** of X means the expected value of $(X - m_1)^n$:

$$n\text{th central moment of } X = \langle (X - m_1)^n \rangle. \tag{2.1.43}$$

The first moment m_1 (or simply m) is called the **mean value** of X, the second central moment is called the **variance** of X and is usually written σ^2, and the square root of the variance is called the **dispersion**. Using (41), we can establish the useful relation

$$\sigma^2 = m_2 - m_1^2 = \langle X^2 \rangle - \langle X \rangle^2. \tag{2.1.44}$$

The following **Tchebycheff inequality** brings out the significance of m and σ: for any $\varkappa > 1$,

$$P(\{|X - m| \geqslant \varkappa\sigma\}) \leqslant 1/\varkappa^2, \tag{2.1.45}$$

i.e. the probability that X differs from its mean by more than $\varkappa\sigma$ is less than $1/\varkappa^2$. *Proof:* If X is continuous, then

$$\sigma^2 = \int (x - m)^2\, \varrho(x)\, dx \geqslant \int_{|x-m| > \varkappa\sigma} (x - m)^2\, \varrho(x)\, dx$$

$$\geqslant \varkappa^2\sigma^2 \int_{|x-m| > \varkappa\sigma} \varrho(x)\, dx = \varkappa^2\sigma^2\, P(\{|X - m| \geqslant \varkappa\sigma\}).$$

The same reasoning goes through if X is discrete.

The mean m is a measure of the location of X since by (45) we are not likely to find X too far from m; for example, the probability of finding X farther than 10σ from m is less than 0.01. Note, however, that m need not be a highly likely value of X; in fact, m might be a point at which the probability distribution is zero. The dispersion σ is a measure of the uncertainty, or statistical spread, in X: if σ is small, then, by (45), X is very likely to be found close to m. A better measure of the spread is the relative dispersion σ/m_1, since quantities measured in the lab are usually of order m_1. Thus, $\sigma/m_1 \ll 1$ means that the spread in X is small on the lab scale.

Example 2 (continued). Defining X by (12), we obtain $m_1 = 3.5$, $m_2 = 15.17$, and using (44), $\sigma^2 = 2.29$. The relative dispersion is $\sigma/m_1 = 0.43$, so the spread in X is rather large. In this example, the mean value $m_1 = 3.5$ is not a possible value of X.

The following **moment theorem** gives an interesting connection between the moments of an r.v. and its probability distribution: the probability distribution of an r.v. is uniquely determined by its moments m_n if they all exist and are such that the series $\Sigma' m_n s^n/n!$ converges absolutely for some number $s \neq 0$. For the proof, see Ref. 1. Thus, in "most" cases (namely, those cases satisfying the conditions of the theorem), the moments of an r.v. uniquely determine its probability distribution. For the mathematical procedure by which we can actually find the probability distribution function from its moments, see Ref. 1.

The extension of these ideas to the case of N r.v. $X, Y, \ldots, Z$ is straightforward: we define expectation values $\langle g(X, Y, \ldots, Z)\rangle$ by analogy with (39) and (40), the moments are $m_{ij\ldots k} = \langle X^i, Y^j \ldots, Z^k\rangle$, and the complete set of moments determines a unique distribution.

Averages of Random Variables. Consider N r.v. $X_1, X_2, \ldots, X_N$ all defined on the same experiment $(\mathscr{S}, \mathbf{F}, P)$. Their average value, defined by

$$\bar{X} = (X_1 + X_2 + \cdots + X_N)/N, \tag{2.1.46}$$

is also an r.v., since $\bar{X}$ is a function $g(X_1, \ldots, X_N)$ of the random variables X_i. If the X_i are all statistically independent of each other, and if they have a common mean m_1 and variance σ_1^2, then the mean and variance of $\bar{X}$ can be shown to be[1]

$$\langle \bar{X}\rangle = m_1 \tag{2.1.47}$$

$$\langle (\bar{X} - m_1)^2\rangle = \sigma_1^2/N. \tag{2.1.48}$$

Thus the average of a set of N statistically independent r.v. X_i (all having the same mean and variance) has the same mean as the X_i, but a variance which is smaller by a factor N^{-1} than the variance of the X_i.

Example 5. A single die is thrown ten times. If we are interested in the average of these ten throws, then we must let the experiment be ten throws of the die, i.e. a *single* trial is ten throws. The "reasonable" (to most people) specification of $(\mathscr{S}, \mathbf{F}, P)$ is as follows: a single outcome is a sequence $\zeta = (f_i^{(1)}, f_j^{(2)}, \ldots, f_k^{(10)})$, where $f_i^{(r)}$ means "face i came up on the rth throw"; $\mathscr{S}$ is the set of all such sequences; $\mathbf{F}$ is all possible collections of such sequences; the probability of any elementary event $\{\zeta\}$ is 6^{-10}, and the probability of any event $\mathscr{E} \in \mathbf{F}$ is the sum of the probabilities of all the different elementary events $\{\zeta\} \subset \mathscr{E}$. Define the ten r.v. $X^{(r)}$ $(r = 1, 2, \ldots, 10)$ by $X^{(r)}(\zeta)$ = number of spots obtained on the

rth throw. The $X^{(r)}$ are statistically independent, with mean 3.50, variance 2.29, and relative dispersion 0.43 (compare Ex. 2, p. 26). The average of ten throws is the r.v.

$$\bar{X}(\zeta) = (1/10)\sum_{r=1}^{10} X^{(r)}(\zeta).$$

By (47) and (48), the mean, variance and relative dispersion of $\bar{X}$ are

$$\langle \bar{X}\rangle = \langle X^{(1)}\rangle = 3.50,$$

$$\langle (\bar{X} - 3.5)^2\rangle = 2.29/10 = 0.229,$$

$$\text{relative dispersion of } \bar{X} = \sqrt{0.229}/3.5 = 0.137.$$

Although the relative dispersion in the outcome of a single throw is large (it is 0.43), the relative dispersion in the average of ten throws is only 0.137. If "relative dispersion" is interpreted as "uncertainty", this says that there is less uncertainty in the average of ten throws than there is in the outcome of one throw, which sounds reasonable.

Equation (48) implies that the variance of the average of N statistically independent r.v. (having common mean and variance) approaches zero as N approaches ∞. This is one version of the so-called **law of large numbers**. If "variance" is interpreted as "uncertainty", this says that the average of a large number of r.v. (having common mean and variance) is almost certain to be equal to the mean of one of the r.v.

2.2 THE INTERPRETATION OF PROBABILITY

One of the nice things about mathematics is that it is clear-cut; it is easy to be clear about objects (such as $\mathscr{S}$, $\mathbf{F}$, and P in Section 2.1) which one just invents. The previous Section was pure mathematics: the section was based on formal assumptions, or axioms, about the abstract objects $\mathscr{S}$, $\mathbf{F}$ and P, and the development consisted in defining new objects (such as random variables) in terms of $(\mathscr{S}, \mathbf{F}, P)$ and deducing relationships between the objects thus defined. Now, this is all very nice as far as it goes. The trouble is, it does not go far enough: the mathematical formalism tells us nothing about the real world. Unfortunately, when we try to learn something about the real world, we find that things become more difficult.

In order to apply probability theory to the real world, we must *interpret* the objects appearing in the theory in terms of human experience. The axiomatic presentation of the previous Section does not constitute such an interpretation. Since the mathematical theory of probability is erected entirely on the postulated properties of $\mathscr{S}$, $\mathbf{F}$, and P, the most economical way to

proceed is to find an interpretation of $\mathscr{S}$, $\mathbf{F}$, and P; the interpretations of all other objects in the formalism will then follow. The problem is that there exist at least four popular interpretations of $(\mathscr{S}, \mathbf{F}, P)$, and innumerable variations within each of the four. Furthermore, it seems that each interpretation contains an irreducible element of obscurity or circularity.

The interpretation of the sample space $\mathscr{S}$ and the field of events $\mathbf{F}$ is reasonably straightforward. The space $\mathscr{S}$ is composed of all those outcomes (or "statements", or "states", or "objects") of interest in a given real world situation. For instance, if we are interested in which apple will be drawn out of a barrel of apples, then the apples are the elements of $\mathscr{S}$; if we are interested in whether Smith will have an automobile accident during the next twelve months, then $\mathscr{S}$ contains the two elements "Smith has an accident" and "Smith does not have an accident"; if we are interested in the state of a mechanical system, then $\mathscr{S}$ consists of all possible states of that system; if we are interested in the sequence of results obtained when a die is thrown ten times, then $\mathscr{S}$ contains one element for each possible sequence of ten throws. An event $\mathscr{E}$ is then a set of such outcomes, and $\mathbf{F}$ is the collection of all such sets $\mathscr{E}$.

The rub comes when we try to assign a real world meaning to the probabilities P. What does "the probability $P(\mathscr{E})$ of an event $\mathscr{E}$" mean? There appear to be at least four different answers, which we will classify as:

1. The empirical, or frequency, interpretation.
2. The probable inference, or inductive logic, interpretation.
3. The subjective, or degree of belief, interpretation.
4. The classical interpretation.

It appears to the author that the second interpretation is the most useful in statistical mechanics, although there is much to be said for the first interpretation. The third interpretation is probably not useful in the physical sciences, and the fourth is just a special case of each of the other interpretations. In the author's opinion, none of these four interpretations is either correct or incorrect; rather, each interpretation is applicable (or correct, or useful) in one context or another. Arguments which purport to show the correctness or incorrectness of a certain interpretation are analogous to arguing that Newton's second law is the "correct" interpretation of second order differential equations, whereas circuit theory is an "incorrect" interpretation.

We will now discuss the four interpretations.

The **frequency interpretation of probability** identifies the probability of an event with the fractional number of times the event occurs in an infinite series

of separate trials. That is,

$$P(\mathscr{E}) = \lim_{n\to\infty} \frac{n(\mathscr{E})}{n}, \tag{2.2.1}$$

where $n(\mathscr{E})$ is the number of times $\mathscr{E}$ occurs in n trials of the experiment under consideration. This interpretation gets its name from the fact that $n(\mathscr{E})/n$ is called the "relative frequency of occurrence of $\mathscr{E}$ during n trials". This point of view was first formulated by Venn,[3] and is supported and discussed in depth by the physicists Von Mises[4] and Reichenbach.[5]

The frequency interpretation is generally considered to apply to situations in which it is possible to carry out a large number of random trials of the experiment under consideration. The precise meaning of "random", when it is used in this context, is not quite clear, but the word means roughly that the sequence of outcomes must not show any regular pattern.

Suppose, for instance, that we are interested in whether or not a throw of a single die will yield an odd number of spots. The event $\mathscr{E}$ of interest is then "odd", and (according to the frequency interpretation) $P(\mathscr{E})$ is simply an abbreviation for the relative number of times the result "odd" comes up in an infinite number of "random" trials. As an illustration of the word "random" in this context, each of the following sequences of outcomes represents trials which are *not* random:

$$(1, 2, 1, 2, 1, 2, \ldots),$$

$$(1, 2, 3, 4, 5, 6, 1, 2, 3, 4, 5, 6, \ldots).$$

As another example, consider an insurance company executive who is interested in whether or not a man, chosen at random from all U.S. males between the ages of 30 and 35, will die in the next 12 months. Here, the experiment consists of choosing at random one person from the set of U.S. males between 30 and 35 years of age. The event of interest is "the man dies in the next 12 months". The probability of this event is then the relative frequency (1), where n is the number of men chosen at random and $n(\mathscr{E})$ is the number of these men who die in the next 12 months. As an example of the word "random", the selection procedure would not be considered random if each of the men selected had epilepsy.

If the experiment under consideration is such that a large number of random trials of the experiment can be carried out, then it is possible (according to the frequency interpretation) to make an *empirical measurement* of $P(\mathscr{E})$ by simply repeating the experiment a large number of times (say n times). The

ratio $n(\mathcal{E})/n$ is then regarded as an experimental measurement of the theoretical result (1). As is always the case in science, the theoretical result is an idealization (in this case the idealization assumes $n \to \infty$) of the actual experimental situation.

One motivation for the frequency interpretation is the desire to give an *empirical* meaning to probabilities, so that $P(\mathcal{E})$ becomes an objective, observable property of the system under consideration. For example, according to the frequency viewpoint, P (odd) is an experimental property of any die, just as the volume and mass are properties of the die. In order to make a measurement of the property P (odd), one simply performs a large number of random throws, and notes the value of n (odd)/n. The measurement becomes more and more precise as n is chosen larger and larger.

The main logical difficulty with the frequency interpretation appears to lie in the condition of randomness. This condition is supposed to mean that the trials are independent in some sense. But in precisely what sense? The statement that the sequence of outcomes must not show any pattern doesn't tell us anything until we know exactly what is and what is not a pattern. If we interpret "randomness" as "statistical independence of successive trials", then we are being circular: statistical independence is meaningless until probability has been defined. If we regard "randomness" to mean that it is impossible to *predict* the outcomes of future trials from the results of past trials, then we have gotten away from a purely empirical view of probability, and the frequency interpretation becomes just a special case of the inductive logic interpretation (to be presented below).

In addition to the above logical problem, there are two severe restrictions on the applicability of the frequency interpretation: (a) this interpretation applies only to situations in which it is possible to carry out a large number of trials; (b) even if a large number of trials is possible, the frequency interpretation says nothing about *individual* trials.

As an example of restriction (a), the frequency interpretation does not apply to such questions as "will there be a nuclear war?" Since there is only one history of the human race, it is unfortunately not possible to carry out more than one trial relative to this question.

As an example of restriction (b), suppose that a large number of trials with a particular die has led to the conclusion that, for this die, P (odd) = 0.6 (apparently this is an unusual die). According to the frequency interpretation, P (odd) = 0.6 is just an abbreviation for the statement that "in a long sequence of random trials, 60% of the throws will be odd". This statement in itself says *nothing* about what will happen on the next throw. That is, according to the frequency interpretation, probability theory merely relates various properties of frequency ratios in long sequences of trials, and says nothing about the outcomes of individual trials. To make the situation more dramatic, suppose that P (odd) = 0.999. According to the frequency interpretation, this simply says that 99.9% of the throws will be odd, and says nothing about what will happen on the next throw. If we are asked to predict the outcome of the next throw, given only that P (odd) = 0.999, and if we hold a strict frequency interpretation of probability, then we can only answer "the throw will be either even or odd; if a more precise prediction is desired, I will need to know the mechanical initial conditions of the system so that I can solve Newton's second law and thus predict the final state". Again, a pure frequentist would be entirely uninfluenced in his smoking habits by statistical data showing that P (cancer | heavy smoking habits) $\gg$ P (cancer | abstinence from tobacco), since the frequentist argues that probabilities are not applicable to individual trials. It appears from this example that pure frequentists may have a shortened life expectancy.

According to the **probable inference interpretation,** or **inductive logic interpretation**, probability theory is the formal expression of inductive reasoning. This point of view was first formulated by the economist Keynes,[6] and is presented in detail by Jeffries[7,8] and Cox.[9] Jaynes[10] has recently made an important contribution to this viewpoint, and a connection between this viewpoint and the frequency interpretation.

In *deductive* logic, we derive consequences as logical implications from statements which are asserted, or postulated, to be true. For instance, we postulate Newton's laws and then deduce the consequence that a projectile acted on only by a constant gravitational field undergoes parabolic motion.

In the case of *inductive* logic, we begin with a general statement (suggested by experience) which seems reasonable but which is not asserted to be true, and we then support the validity of, or partially verify, this statement by establishing (experimentally) the truth of some of its implications. For example, we partially verify Newton's laws each time we observe a projectile undergoing parabolic motion in a constant gravitational field. We cannot *prove* (i.e. deduce) Newton's laws from such observations, since the parabolic motion of a projectile or of any number of projectiles does *not* logically imply Newton's laws. But our confidence in Newton's laws is increased by such observations, and thus we say that we "induce" (not "deduce") Newton's laws from the observations. Clearly, induction is just a high-class way of guessing. Much scientific reasoning is of this inductive type; in fact, we establish and support all basic scientific laws only by inductive reasoning.

The inductive logic viewpoint of probability holds that probability theory is simply the formalism for this inductive mode of reasoning.

In deductive logic, every statement is either true or false. Thus, to indicate the status of a statement q, we need only two symbols: one for "q is true" and one for "q is false". In inductive logic, statements may be true on the basis of the given data, or false on the basis of the given data, or somewhere between these two extremes. That is, a statement q may be entitled to only partial assent on the basis of the given data. Thus, the formalism of inductive logic must allow the status of a given statement to range over a continuum of truth values between "q is true" and "q is false". According to the inductive logic viewpoint of probability theory, $P(q)$ is just the quantity which accomplishes this. That is, this viewpoint holds that, given any data D relative to some experiment under consideration, and given any event $\mathcal{E}$ relative to the same experiment, there exists a unique probability $P(\mathcal{E})$ which represents the inductive logical status of $\mathcal{E}$ on the basis of D. The extreme case $P(\mathcal{E}) = 1$ means "$\mathcal{E}$ is certain on the basis of D", in other words, "D implies $\mathcal{E}$". The other extreme, $P(\mathcal{E}) = 0$, means "D implies that $\mathcal{E}$ is false". Intermediate values, $0 < P(\mathcal{E}) < 1$, then apply to the case (which is the case usually encountered in scientific inductive reasoning) in which D does not imply either $\mathcal{E}$ or its negation.

According to the inductive logic viewpoint, the data D uniquely determines the probability assignment P: for a given sample space $\mathcal{S}$ and field $\mathbf{F}$, and for given data D relative to $(\mathcal{S}, \mathbf{F})$, every observer will assign the same P. We will give the actual method of determining P in Section 2.4. Hence, the inductive logic viewpoint is not a subjective (i.e. relative to the observer) viewpoint. On the other hand, this viewpoint is not purely objective (i.e. relative only to the object), since probabilities depend on the given data.

For example, in the experiment of throwing a single die, P (even) $= 1/2$ is not regarded as a property of the die itself, but is instead regarded as a property of the die and of the given information. If the information is that the die has 6 sides, with i dots on the ith side, then the appropriate assignment is P (even) $= 1/2$, while if the information is the above statement in addition to the information that the die is weighted in such a way that only $i = 1, 2,$ or 3 can come up, then the appropriate assignment is P (even) $= 1/3$. Two observers will assign different probabilities to the die experiment if (and only if) the given data is different for the two observers. As a more dramatic example, consider two observers who wish to make bets on the next outcome of a die experiment. One observer knows only that the die is an ordinary six-sided die, while the other knows the exact mechanical initial conditions

(initial mechanical state of the die, air resistance, elastic properties of the die and table, etc.) and is able to solve for the exact motion and can thus predict that the next outcome will be, say, 4. The first observer assigns the probabilities $P(\{f_i\}) = 1/6$ $(i = 1, 2, \ldots, 6)$ to the next trial, while the second observer assigns $P(\{f_4\}) = 1$, $P(\{f_i\}) = 0$ $(i \neq 4)$.

The main drawback (if we choose to consider it a drawback) to the inductive logic interpretation is that the "probability of $\mathcal{E}$ given the data D" is not an experimentally measured number, but is simply a statement about the logical status (or "reasonableness") of $\mathcal{E}$. Thus, it appears that the word *probability* remains a primitive undefined term in the inductive logic interpretation. This is not surprising, since in the formalism of deductive logic the truth values *true* and *false* are also undefined.

The inductive logic viewpoint applies to a broader class of situations than does the frequency viewpoint. We can assign inductive probabilities even though the trials are not random, and even though a large number of trials is impossible. Furthermore, an inductive probability is not merely a description of a frequency ratio over a long series of trials, but is instead an assertion about a *single* trial. For instance, P (odd) $= 0.999$ says not only that it is expected that 99.9% of the throws in a long sequence will be odd, but also that it is highly likely, or quite reasonable, that the *next* throw will be odd. An inductivist *will* be influenced in his smoking habits by statistical data on smoking and health.

To summarize: the inductive logic viewpoint holds that probability theory is the formalism to which we resort when we lack sufficient data to make deductive inferences. Probability theory becomes the formalism for describing situations in which the data does not imply a definite outcome. The probabilistic description is objective in the sense that it depends not on the observer but only on the experiment and on the observer's data.

We will describe the remaining viewpoints more briefly, since they do not seem to be of significance in the physical sciences.

According to the **subjective interpretation,** probabilities represent degrees of belief, so that $P(\mathcal{E})$ is relative not only to the experiment and to the given data but also the inclination of the person whose degree of belief that probability represents. Probability theory then becomes the scheme for establishing consistent relationships between different degrees of belief. For example, given a die and data D, observer A might assign P (odd) $= 0.4$, and observer B (who happens to feel more favorably inclined toward "odd") might assign P (odd) $= 0.7$. The theory of probability then obligates A to assign P(even) $= 0.6$, and B to assign P(even) $= 0.3$. Several outstanding mathematicians

(e.g. E. Borel and B. Koopman) hold this viewpoint. Reference 11 contains a collection of fundamental essays in subjective probability theory.

The **classical interpretation** is not really a separate interpretation; we include it here mainly for historical completeness. The classical interpretation seems to have been the only explicitly formulated interpretation from the time of the first systematic use of probability theory around 1650 (in connection with the widespread infatuation with games of chance in 17th century France) until the first formulation, by Venn in 1886, of the frequency interpretation. The classical interpretation holds that probability theory applies only to situations in which the set $\mathscr{S}$ of outcomes ζ has sufficient symmetry that no grounds exist for preferring one outcome over another. The probability $P(\mathscr{E})$ is then defined as the ratio of the number of outcomes in the event $\mathscr{E}$ to the number of outcomes in $\mathscr{S}$:

$$P(\mathscr{E}) = \frac{\text{number of } \zeta \text{ such that } \zeta \in \mathscr{E}}{\text{number of } \zeta \in \mathscr{S}}. \tag{2.2.2}$$

For example, in the case of the die, the symmetry between the different possible outcomes 1, 2, ..., 6 seems to leave no grounds for preferring one outcome over another. Thus, it seems reasonable to assign equal probabilities to each side, and hence the probability (2) to an arbitrary event $\mathscr{E}$.

We can consider this interpretation as merely a special case of any of the first three interpretations, where one establishes symmetry by different means. For example, in the case of the die experiment, if a long sequence of throws results in a relative frequency of 1/6 for each side, then the frequency interpretation leads to (2); if the given data contains no evidence favoring any outcome over any other, then the inductive logic interpretation leads to (2); if the observer feels just as favorably inclined toward any one outcome as toward any other, then the subjective interpretation leads to (2).

At this point, it might be a good exercise for the reader to go back over the formalism presented in Section 2.1, interpreting the expressions and concepts in terms of the relative frequency and inductive logic viewpoints. For example, the frequency viewpoint interprets (2.1.11) (the conduction for statistical independence of $\mathscr{E}_1$ and $\mathscr{E}_2$) as saying that the relative frequency of $\mathscr{E}_1$ in a long sequence of random trials is the same as the relative frequency of $\mathscr{E}_1$ in the sub-sequence consisting of outcomes in which $\mathscr{E}_2$ occurred; the inductive logic viewpoint interprets (2.1.11) as implying that the data "$\mathscr{E}_2$ occurred" is irrelevant to predictions made about $\mathscr{E}_1$.

According to the viewpoint adopted in Chapter 1, statistical mechanics is the study of mechanical systems on the basis of incomplete data. Predictions

about incompletely specified systems cannot in general be *deductive* predictions; for instance, it is impossible to predict with certainty the precise number of molecules in a given sub-volume of a container, given only the total energy, total volume, and total number of molecules in the container. Statistical mechanical predictions contain an element of uncertainty, and are based on *inductive* reasoning. Since the formalism for inductive reasoning is probability theory, it follows that probability theory forms the mathematical backbone of statistical mechanics. Thus, as far as statistical mechanics is concerned, the natural (i.e. useful) interpretation of probability is the inductive logic interpretation. "Probability" will be understood in this sense throughout the remainder of this book.

2.3 INFORMATION

Suppose you are given a set $\mathscr{E}$ of possible outcomes ζ in some physical experiment. Initially, you do not know which outcome will occur. Now a friend tells you "the outcome is ζ_1". Your friend has given you some information; it would be useful to have a quantitative measure of this information. In this Section, we will discuss such an information measure.

Information theory[12–16] is the quantitative study of the acquisition, production and transmission of information. The theory was founded in 1948 when C. E. Shannon introduced a useful quantitative measure of the missing information in a probability distribution.[12] During the past two decades, information theory has grown into a broad, highly developed body of knowledge, with ramifications in communications engineering (to which the theory was originally applied), psychology, biology, physics, and pure mathematics. In this book, we will discuss only those aspects of information theory which are pertinent to statistical mechanics.

If we accept the view, proposed in Chapter 1, that statistical mechanics is the study of mechanical systems for which the given information is incomplete, then we are naturally inclined to try applying information theory to statistical mechanics. More generally, information theory concepts should be relevant to any field in which inductive probabilities are useful. The reason is that inductive probabilities arise whenever the given information is not sufficient to permit deductive inferences; any theory which purports to study information quantitatively is likely to be useful in such a situation.

Returning to the problem posed in the opening paragraph of this Section, let us put the question in a more quantitative form. Suppose that your original state of knowledge about the outcome is represented by the prob-

ability assignment P^0. For simplicity, assume that $\mathscr{S}$ is finite, so that the probability assignment can be given by the numbers $P^0(\{\zeta_i\}) = p_i^0 \ (i = 1, 2, \ldots, n)$. Your friend now presents new data which changes your state of knowledge to $P(\{\zeta_i\}) = p_i \ (i = 1, 2, \ldots, n)$. (In the example given in the opening paragraph, the new data gave the exact outcome; more generally, the new data might only partially specify the outcome). The problem is to find a quantitative measure for the information provided by the new data.

Thus, we seek a quantity

$$I(P;P^0) = I(p_1, p_2, \ldots, p_n; p_1^0, p_2^0, \ldots, p_n^0) \tag{2.3.1}$$

which represents the information content of the probability assignment $(p_1, \ldots, p_n)$, when the prior probabilities were $(p_1^0, \ldots, p_n^0)$. We will call this quantity the "information in P relative to P^0", or simply the **information**.* Note that the information, as defined here, is dependent on *two* schemes $(\mathscr{S}, \mathbf{F}, P^0)$ and $(\mathscr{S}, \mathbf{F}, P)$ having the same sample space $\mathscr{S}$ and field of events $\mathbf{F}$, but different probability assignments.

The problem of determining a reasonable measure of information content is one of the basic problems of information theory. C. E. Shannon[12] was the first to solve this problem. However, Shannon (and most subsequent authors) stated the problem in a form which was somewhat different from the statement given above; the solution to the problem posed by Shannon was that the measure of the *missing* information in a *single* probability assignment $(p_1, \ldots, p_n)$ is

$$-k \sum p_i \ln p_i \text{**} \tag{2.3.2}$$

where k is a positive constant. Shannon's measure (2) suffers from the defects that it cannot be consistently generalized from discrete to continuous probability spaces, and when it is (non-rigorously) generalized, the resulting expression

$$-k \int \varrho(x) \ln \varrho(x) \, dx \tag{2.3.3}$$

is not invariant under a change of variables $x \to y = y(x)$. Our formulation, on the other hand, will lead to an expression for $I(P; P^0)$ which does not suffer from these defects.

Now, one cannot just conjure up a theory of information out of thin air. If we wish to find an expression for $I(P; P^0)$, we will have to begin by assuming some reasonable properties for $I(P; P^0)$. Proceeding in this way, we will prove a uniqueness theorem of the following form: if we postulate that $I(P; P^0)$ has certain prescribed (and reasonable) properties, then $I(P; P^0)$ necessarily has the explicit form $I(P; P^0) = \cdots$ An alternative (and less satisfactory) way of proceeding would be to simply postulate an explicit mathematical expression for I, and then show that this expression has properties which should be expected of an information measure. The difficulty with this approach is that it would leave us in doubt as to whether the postulated expression is the *only* one having the stated properties. The uniqueness theorem approach will be used here.

* Information theorists often refer to $I(P; P^0)$ as an *entropy*; expression (2) is also often called an *entropy*. We will reserve the word *entropy* for a more specific use in statistical mechanics.

** All logarithms are taken to the base e.

We seek a function of the form (1), defined for any pair of probability assignments P, P^0 on a finite sample space $\mathscr{S}$. It is reasonable to assume that

$$I(p_1, \ldots, p_n; p_1^0, \ldots, p_n^0) \text{ is a continuous function,} \tag{2.3.4}$$

and

$$I(p_1, \ldots, p_j, \ldots, p_k, \ldots, p_n; p_1^0, \ldots, p_j^0, \ldots, p_k^0, \ldots, p_n^0) = I(p_1, \ldots, p_k, \ldots, p_j, \ldots, p_n; p_1^0, \ldots, p_k^0, \ldots, p_j^0, \ldots, p_n^0), \tag{2.3.5}$$

and

$$I(P; P) = 0. \tag{2.3.6}$$

Postulate (4) says that a small change in P and P^0 changes the information by only a small amount; (5) says that the information does not depend on the manner in which the outcomes are labeled; (6) says that no information is obtained if the final probabilities are the same as the prior probabilities.

For any pair of integers n and n_0 such that $n_0 \geqslant n$, the expression $I(1/n, \ldots, 1/n, 0, \ldots, 0; 1/n_0, \ldots, 1/n_0)$ represents the information obtained when the number of equally likely possibilities is reduced from n_0 to n. It is reasonable to assume that

$$I(1/n, \ldots, 1/n, 0, \ldots, 0; 1/n_0, \ldots, 1/n_0) \text{ is an increasing function of } n_0 \text{ and a decreasing function of } n, \text{ for any integers } n \text{ and } n_0 \text{ such that } n_0 \geqslant n. \tag{2.3.7}$$

Postulate (7) says, for instance, that the information obtained upon reducing the number of equally likely sides on a die from 6 to 3 is greater than the information obtained upon a reduction from 6 to 4.

Finally, a condition known as the composition rule will be needed. Suppose that the sample space $\mathscr{S}$ is divided into two sub-spaces $\mathscr{S}_1 = \{\zeta_1, \ldots, \zeta_r\}$

and $\mathscr{S}_2 = \{\zeta_{r+1}, \ldots, \zeta_n\}$. The sub-spaces $\mathscr{S}_1, \mathscr{S}_2$ are events (i.e. collections of outcomes) having final probabilities

$$q_1 = p_1 + \cdots + p_r \quad \text{and} \quad q_2 = p_{r+1} + \cdots + p_n, \tag{2.3.8}$$

and prior probabilities

$$q_1^0 = p_1^0 + \cdots + p_r^0 \quad \text{and} \quad q_2^0 = p_{r+1}^0 + \cdots + p_n^0. \tag{2.3.9}$$

The final probabilities of ζ_i, given that $\zeta_i \in \mathscr{S}_1$ (i.e. given that $\mathscr{S}_1$ occurred—compare (2.1.8)) are $P(\{\zeta_i\}) | \mathscr{S}_1) = p_i/q_1$ $(i = 1, \ldots, r)$, and the prior probabilities of ζ_i, given $\mathscr{S}_1$, are $P^0(\{\zeta_i\} | \mathscr{S}_1) = p_i^0/q_1^0$ $(i = 1, \ldots, r)$. Similarly, the final and prior probabilities of ζ_i given $\mathscr{S}_2$ are p_i/q_2 $(i = r+1, \ldots, n)$ and p_i^0/q_2^0 $(i = r+1, \ldots, n)$ respectively. Now, we may give information about the outcome *either* by specifying the probabilities $p_1, \ldots, p_n$ directly, *or* by specifying the probabilities q_1, q_2 of the sub-spaces $\mathscr{S}_1, \mathscr{S}_2$ and then giving the conditional probabilities p_i/q_1, p_i/q_2 within these sub-spaces. These two alternatives are shown diagramatically in Figures 1 and 2. We now postulate

$$\begin{aligned} &I(p_1, \ldots, p_r, p_{r+1}, \ldots, p_n; p_1^0, \ldots, p_r^0, p_{r+1}^0, \ldots, p_n^0) \\ &= I(q_1, q_2; q_1^0, q_2^0) + q_1 I\left(\frac{p_1}{q_1}, \ldots, \frac{p_r}{q_1}; \frac{p_1^0}{q_1^0}, \ldots, \frac{p_r^0}{q_1^0}\right) \\ &\quad + q_2 I\left(\frac{p_{r+1}}{q_2}, \ldots, \frac{p_n}{q_2}; \frac{p_{r+1}^0}{q_2^0}, \ldots, \frac{p_n^0}{q_2^0}\right). \end{aligned} \tag{2.3.10}$$

Postulate (10) asserts that the amount of information in the scheme of Figure 1 equals the amount of information in the equivalent scheme of Figure 2, and that furthermore the right hand side of (10) is a reasonable expression for the information in the scheme of Figure 2. Thus, the basic idea of the composition rule is the assumption that the information in the scheme of Figure 2 is given by the *sum* of the information $I(q_1, q_2; q_1^0, q_2^0)$ concerning the first step (i.e. concerning which sub-space occurred), plus the weighted

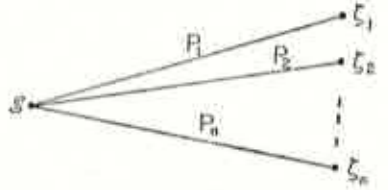

Figure 2.3-1 Diagram of the probability assignment $(p_1, \ldots, p_n)$

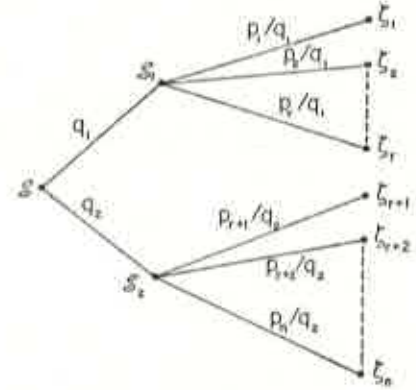

Figure 2.3-2 An alternative way of describing the probability assignment shown in Figure 1

(with weighting factors q_1 and q_2) information concerning the second step. This seems reasonable if you think about it awhile; at least, the right hand side of (10) seems more reasonable than any *other* choice of an expression for the information in the scheme of Figure 2.

Uniqueness Theorem. Assume that $I(p_1, \ldots, p_n; p_1^0, \ldots, p_n^0)$ exists for any pair of probability assignments P, P^0 on a finite sample space $\mathscr{S} = \{\zeta_1, \ldots, \zeta_n\}$, and assume that I satisfies the postulates (4), (5), (6), (7), and (10). Then I is necessarily given by

$$I(p_1, \ldots, p_n; p_1^0, \ldots, p_n^0) = k \Sigma p_i \ln (p_i/p_i^0),$$

where k is a positive constant. (2.3.11)

The proof is in Appendix A. Thus, (11) is the *only* expression having the intuitively reasonable properties (4), (5), (6), (7), and (10). For simplicity, we will henceforth choose $k = 1$.

Expression (11) applies to the case of a discrete, finite sample space $\mathscr{S}$. In case $\mathscr{S}$ is discrete but infinite, the obvious generalization of (11) is

$$I(P; P^0) = \sum_{i=1}^{\infty} p_i \ln (p_i/p_i^0), \tag{2.3.12}$$

i.e. the information in an infinite scheme $(p_1, p_2, \ldots)$ is defined as the limit (as $n \to \infty$) of the information in the finite scheme $(p_1, \ldots, p_n)$.

It is not difficult to find a reasonable generalization of (11) for the case that the probability assignments P and P^0 are both continuous. Let $\mathscr{S}$ be the segment $\{a \leqslant x \leqslant b\}$ of the real axis, let $\mathbf{F}$ be the set of all Lebesgue measurable* subsets of $\mathscr{S}$ and assume that P and P^0 are continuous probability assignments with probability densities $\varrho(x)$, $\varrho^0(x)$.**

Divide $\mathscr{S}$ into segments by means of the partition or "net" $(x_0 = a, x_1, \ldots, x_n = b)$, and let $\Delta x_i = x_i - x_{i-1}$ and $\mathscr{E}_i = \{x_{i-1} < x \leqslant x_i\}$. The probabilities

$$P(\mathscr{E}_i) = \int_{\mathscr{E}_i} \varrho(x)\,dx \simeq \varrho(x_i)\,\Delta x_i$$
$$P^0(\mathscr{E}_i) = \int_{\mathscr{E}_i} \varrho^0(x)\,dx \simeq \varrho^0(x_i)\,\Delta x_i \tag{2.3.13}$$

then form a discrete set, with information

$$I_{\text{net}} = \Sigma\, P(\mathscr{E}_i) \ln [P(\mathscr{E}_i)/P^0(\mathscr{E}_i)] \simeq \Sigma\, \varrho(x_i)\,\Delta x_i \ln [\varrho(x_i)/\varrho^0(x_i)].$$

As the net becomes finer, the approximations become equalities, and I_{net} approaches the limiting value

$$I(P; P^0) = I[\varrho(x); \varrho^0(x)] = \int_a^b \varrho(x) \ln \left[\frac{\varrho(x)}{\varrho^0(x)}\right] dx. \tag{2.3.14}$$

From the manner in which we obtained (14), it is reasonable to take (14) as the information in the probability assignment $\varrho(x)$, relative to the prior probability assignment $\varrho^0(x)$. Equation (14) may be generalized in an obvious way to the case that $\mathscr{S}$ is an N-dimensional continuum.

We will now present a few simple properties of the information measures (11) and (14); hopefully, these properties will further convince the reader of the reasonableness of (11) and (14).

> For either the discrete or continuous case, $I(P; P^0) \geqslant 0$, with equality if and only if P and P^0 are identical. (2.3.15)

Proof: For any positive number y, $\ln y \geqslant 1 - y^{-1}$ with equality only at $y = 1$. Letting $y = p_i/p_i^0$, we obtain

$$\Sigma\, p_i \ln (p_i/p_i^0) \geqslant \Sigma\, (p_i - p_i^0) = 0$$

* See footnote page 18.

** It may be worth noting, for the sake of mathematical clarity, that *probability density* is used here in a slightly different sense from that given in Section 2.1. In Section 2.1, we defined $\varrho(x)$ on the range of some r.v. X. Here, we define $\varrho(x)$ directly on the sample space $\mathscr{S}$.

with equality only if P and P^0 are identical. For the continuous case, set $y = \varrho(x)/\varrho^0(x)$. Property (15) says that we obtain information whenever we learn something new, i.e. whenever the final state of knowledge P differs from the initial state P^0.

> In the discrete case, if $p_i = \delta_{ik}$ $(i = 1, \ldots, n;\ k$ fixed) then $I(P; P^0) = \ln (1/p_k^0)$. (2.3.16)

Here, δ_{ik} is the Kronecker delta (0 if $i \neq k$, 1 if $i = k$). In deriving (16), $p_i \ln (p_i/p_i^0)$ at $p_i = 0$ is defined by

$$p_i \ln \left(\frac{p_i}{p_i^0}\right)\bigg|_{p_i=0} \equiv \lim_{p_i \to 0} \left\{p_i \ln \left(\frac{p_i}{p_i^0}\right)\right\} = 0. \tag{2.3.17}$$

Property (16) says that the information content of the statement "ζ_k occurred" depends upon how strongly the prior data P^0 indicated the occurrence of ζ_k. If P^0 *implied* ζ_k (i.e. if $p_k^0 = 1$), then the new statement contains nothing new and hence I is zero. If P^0 assigns only a small probability $p_k^0 \ll 1$ to ζ_k, then the statement is very informative and hence I is large.

> In the discrete case, if P^0 is given by a set of n_0 equally likely outcomes, and P is given by a subset of n equally likely outcomes, then $I(P; P^0) = \ln (n_0/n)$. (2.3.18)

Property (18), which follows by direct substitution into (11), says that any statement which reduces the number of equally probable outcomes from n_0 to n contains $\ln (n_0/n)$ units of information. Larger reductions correspond to larger amounts of information. A similar result holds for the continuous case.

> In the discrete case, if (for any i) $p_i^0 = 0$ and $p_i \neq 0$, then $I(P; P^0)$ is infinite. (2.3.19)

The infinity arises from the term $p_i \ln (p_i/0)$ in (11). Thus we obtain an infinite amount of information upon being told that an event $\{\zeta_i\}$ which had been thought impossible is actually possible. From another point of view, (19) states that I is undefined whenever the new data is inconsistent with the prior data.

> If P^0 is continuous and P is discrete, then $I(P; P^0)$ is infinite. (2.3.20)

This result follows from substituting $\varrho(x) = \Sigma p_i \, \delta(x - x_i)$ into (14). Property (20) implies that an infinite amount of information is needed to pick a single point (or a countable set of points) out of a continuum.

> For continuous probability assignments P and P^0, $I(P; P^0)$ is invariant under a monotone one-to-one transformation $x \to y = g(x)$. (2.3.21)

More precisely, letting P and P^0 be described by the densities $\varrho(x)$ and $\varrho^0(x)$ in terms of x, and by $\bar{\varrho}(y)$ and $\bar{\varrho}^0(y)$ in terms of $y = g(x)$, and defining

$$I_x(P; P^0) \equiv \int_a^b \varrho(x) \ln \left[\frac{\varrho(x)}{\varrho^0(x)} \right] dx,$$

$$I_y(P; P^0) \equiv \int_{g(a)}^{g(b)} \bar{\varrho}(y) \ln \left[\frac{\bar{\varrho}(y)}{\bar{\varrho}^0(y)} \right] dy,$$

(21) asserts that $I_x(P; P^0) = I_y(P; P^0)$. The proof will be left to the reader. This property insures that the information in the probability assignment P, relative to P^0, does not depend on the variable chosen to describe the sample space $\mathscr{S}$. That is, the information content of a message is independent of the "language" in which the message is given.

We will now use the basic concept of the information $I(P; P^0)$ to develop the concepts of the **relevance** $R(X, Y)$ of the r.v. Y to the r.v. X, and the **missing information** (or **uncertainty**) $U(P)$ in a probability assignment P.

The **relevance** $R(X, Y)$ of Y to X means the information in the joint distribution of X and Y, when the prior distribution is the product of the distribution for X and the distribution for Y. Thus, for discrete r.v.,

$$R(X, Y) = \sum_{ij} P_{XY}(x_i, y_j) \ln \left[\frac{P_{XY}(x_i, y_j)}{P_X(x_i) P_Y(y_j)} \right]. \tag{2.3.22}$$

For continuous r.v.,

$$R(X, Y) = \int\int \varrho_{XY}(x, y) \ln \left[\frac{\varrho_{XY}(x, y)}{\varrho_X(x) \varrho_Y(y)} \right] dx \, dy. \tag{2.3.23}$$

We will give a few mathematical properties of $R(X, Y)$. From (15) and the definition of statistical independence (see Section 2.1),

> $R(X, Y) \geqslant 0$ with equality if and only if X and Y are statistically independent. (2.3.24)

Since (22) and (23) are symmetric in X and Y,

$$R(X, Y) = R(Y, X). \tag{2.3.25}$$

From (22), noting that $p_{XX}(x_i, x_j) = p_X(x_i) \, \delta_{ij}$, we have:

> The relevance of any discrete r.v. X to itself is
>
> $$R(X, X) = -\Sigma p_X(x_i) \ln p_X(x_i). \tag{2.3.26}$$

By either generalizing (26) to the case of a continuous r.v. X, or by setting $\varrho_{XX}(x, x') = \varrho_X(x) \, \delta(x - x')$ in (23), we have:

> The relevance of any continuous r.v. X to itself is
>
> $$R(X, X) = \infty. \tag{2.3.27}$$

The relevance $R(X, Y)$ measures the information contained in the joint distribution for X and Y, if one already knows the individual distributions for X and Y alone. Now, the individual distributions contain the same information as the joint distribution concerning X and Y separately; however, the individual distributions do not tell us how to readjust our predictions about X, if new information is given about Y. Thus, R measures the correlational information in the joint distribution, i.e. the amount of information (contained in the joint distribution) about X which would be gained by knowledge of Y. Property (24) states that we can never lose information about X by gaining knowledge about Y, and that we in fact gain information about X (by learning about Y) whenever X and Y are correlated. Equation (25) states that knowledge of Y provides the same amount of information concerning X as knowledge of X does concerning Y. Expression (26) for the relevance of X to itself is identical with Shannon's measure of the missing information, equation (2). The relevance $R(X, X)$ of X to itself means the information gained about X upon learning the value of X, when the prior distribution is $p_X(x_i)$. Thus it is reasonable to call $R(X, X)$ the missing information in $p_X(x_i)$. However, our measure of the missing information, to be given below, will be somewhat different from $R(X, X)$ (our measure will be a generalization of (26)). Property (27) shows that when we generalize Shannon's information measure (26) to continuous distributions, the result diverges.

We now come to the important concept of the uncertainty, or missing information, in a probability assignment. Suppose that the probability assignment $P^m = (p_1^m, p_2^m, \ldots, p_n^m)$ corresponds to the maximum knowledge which can be obtained about the outcome ξ. The information in P^m, relative

to the prior probabilities P^0, is then $I(P^m; P^0)$; thus, $I(P^m; P^0)$ is the maximum information obtainable, relative to the prior probabilities P^0. Now, if one's actual state of knowledge is described by the probability assignment P, then the missing information needed to attain the state P^m is

$$U(P; P^m; P^0) = I(P^m; P^0) - I(P; P^0). \tag{2.3.28}$$

Equation (28) will be our official definition of the **missing information**, or **uncertainty**, in the probability assignment P. The uncertainty depends, as does the information, on the prior probabilities P^0; it also depends on the probabilities P^m which are taken as describing the state of maximum knowledge. We will often denote the uncertainty by $U(P)$, omitting explicit mention of P^m and P^0.

We will give a few formal properties of $U(P)$. Equations (20) and (28) imply:

> If P and P^0 are continuous while P^m is discrete, then $U(P)$ is infinite. (2.3.29)

Equations (11) and (28) imply:

> Let the sample space be the finite set $\zeta_1, \zeta_2, \ldots, \zeta_n$, and assume that P^0 and P^m are given by
>
> $$P^0(\{\zeta_i\}) = 1/n \quad (i = 1, \ldots, n),$$
>
> $$P^m(\{\zeta_i\}) = \delta_{ik} \quad (i = 1, \ldots, n;\ k \text{ fixed});$$
>
> then
>
> $$U(P) = U(p_1, \ldots, p_n) = -\Sigma\, p_i \ln p_i. \tag{2.3.30}$$

Similarly, (14) and (28) imply:

> Let the sample space be the region $(a \leqslant x \leqslant b)$ of the real axis. Assume that P is continuous, and let P^0 and P^m be given by $\varrho^0(x) = \text{const.} = (b-a)^{-1}$,
>
> $$\varrho^m(x) = \begin{cases} L^{-1} & (x_1 \leqslant x \leqslant x_1 + L) \\ 0 & (x < x_1 \text{ or } x > x_1 + L). \end{cases} \tag{2.3.31}$$
>
> Then
>
> $$U(P) = U[\varrho(x)] = -\int \varrho(x) \ln [L\varrho(x)]\, dx.$$

Property (31) has an obvious generalization to the case that the sample space is an N-dimensional continuum.

Property (29) states that the uncertainty in a continuous distribution $\varrho(x)$, when the state of maximum knowledge is regarded as being given by a single point x_0 (or a countable set of points), is infinity. Thus, the uncertainty (28) is not a very useful concept if P is continuous while P^m is discrete. Properties (30) and (31) give the expressions for $U(P)$ in two cases which occur frequently in applications. Comparing (30) with (2), we see that $U(P)$ reduces to Shannon's information measure in the special case referred to in (30). Comparing (31) with (3), we see that $U(P)$ is similar to (but not identical with) Shannon's measure in the special case referred to in (31).

A few examples will help illustrate the nature of the uncertainty measures (30) and (31). In each example, we assume that the reference states P^0 and P^m satisfy the conditions of (30) and (31).

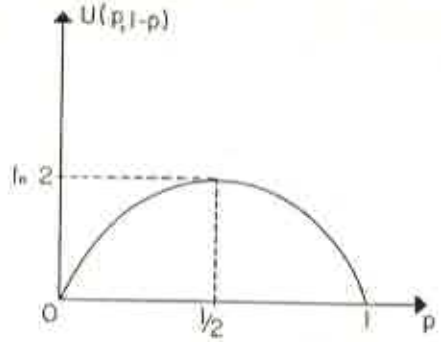

Figure 2.3-3 The uncertainty in a two-sided experiment, with probabilities p and $1-p$.

The uncertainty in the outcome of flipping a well-balanced coin is $U(1/2, 1/2) = \ln 2$. More generally, the uncertainty in any experiment having two outcomes, with $P(\{\zeta_1\}) = p$ and $P(\{\zeta_2\}) = 1 - p$, is

$$U(p, 1-p) = -p \ln p - (1-p) \ln (1-p).$$

This function is sketched in Figure 3. As expected, the uncertainty is maximum when $p = 1/2$, and decreases to zero as p approches 0 or 1.

The uncertainty in any experiment having n equally likely outcomes is $U(1/n, 1/n, \ldots, 1/n) = \ln n$. As expected, the uncertainty increases with n.

The uncertainty in a Gaussian distribution, i.e. a continuous distribution with density

$$\varrho(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-(x-m)^2}{2\sigma^2}\right], \tag{2.3.32}$$

may be found as follows: We can show from (32) that the parameters m and σ^2 are the mean and variance of $\varrho(x)$. From (31) (upon extending the limits to $(-\infty, \infty)$),

$$U[\varrho(x)] = -\int_{-\infty}^{\infty} \varrho(x) \ln \left\{ \frac{L}{\sigma\sqrt{2\pi}} \exp\left[\frac{-(x-m)^2}{2\sigma^2} \right] \right\} dx$$

$$= -\left\langle \ln\left(\frac{L}{\sigma\sqrt{2\pi}} \right) \right\rangle - \left\langle \frac{-(x-m)^2}{2\sigma^2} \right\rangle$$

$$= -\ln\left(\frac{L}{\sigma\sqrt{2\pi}} \right) + \frac{1}{2\sigma^2} \langle (x-m)^2 \rangle = -\ln\left(\frac{L}{\sigma\sqrt{2\pi}} \right) + \frac{1}{2}$$

$$= \ln\left(\sqrt{2\pi e}\, \frac{\sigma}{L} \right). \tag{2.3.33}$$

As expected, U increases with σ, decreases as L increases, and is independent of m.

2.4 JAYNES' PRINCIPLE

In this Section, we give an important principle of inductive reasoning first stated by E. T. Jaynes in 1957 in connection with statistical mechanics[17], and used by Jaynes[10] and others[16,18] in more general contexts.

Suppose that you must make a prediction about the outcome of some experiment, but that the given data is incomplete (i.e. does not determine a precise outcome). What should you do? One answer is that you should do nothing until someone hands you more data. But suppose that you must make a prediction *without* having complete data. In the real world, nearly all predictions are of this type. For example, you may want to make predictions about a box of gas on the basis of measured values of only a few parameters such as the energy, volume, and number of moles. It would not be very useful in this situation to wait until all the data (i.e. the precise positions and velocities of all the particles!) is known before making predictions. Thus, a more practical answer to our question is that you should *induce* (i.e. describe probabilistically) the outcome from the data. Thus our problem is to find a probability assignment P corresponding to the given data.

For example, suppose that we want to predict the number of spots showing on the next trial in the die-throwing experiment, given only that the die has six sides with i spots on the ith side, and that the average number of spots

obtained in a previous long series of throws was 4.5 (this is an unusual die)[19]. The data is incomplete, so we must use probabilistic reasoning. That is, we seek a probability assignment $P(\{f_i\}) = p_i\ (i = 1, 2, \ldots, 6)$ representing the most reasonable statement which can be made about the outcome of the next trial. Since the p_i are probabilities,

$$\Sigma\, p_i = 1. \tag{2.4.1}$$

Since the average throw is 4.5, the p_i should satisfy

$$\langle i \rangle = \Sigma\, i p_i = 4.5.^{*} \tag{2.4.2}$$

Many different probability assignments agree with (1) and (2). One such assignment is $p_4 = 1/2$, $p_5 = 1/2$, $p_1 = p_2 = p_3 = p_6 = 0$. But this assignment seems to assume arbitrarily that f_1, f_2, f_3, f_6 cannot occur, whereas the data does not imply this. That is, this assignment contains more information than is actually given by the data. What we would like to find is that probability assignment which contains *only* the information contained in the data. This is in accord with the standard scientific procedure of assuming no more than is given experimentally. But we already have a quantitative expression $I(P; P^0)$ for the information in a probability assignment P. Suppose that we *minimize* $I(P; P^0)$, for fixed P^0, with respect to those probability assignments P satisfying the given data, i.e. satisfying (1) and (2). The minimizing probability assignment $P^{\min}$ then contains all the information contained in the data, since $P^{\min}$ implies (1) and (2). Furthermore, any other set P^1 satisfying (1) and (2) but *not* minimizing $I(P; P^0)$ also contains the information in the data but contains, in addition, an amount of information $I(P^1; P^0) - I(P^{\min}; P^0)$. That is, P^1 contains more information than is given by the data. Thus the unique probability assignment, which contains the given data and does not arbitrarily assume anything *more* than the given data, is that assignment which minimizes $I(P; P^0)$, for fixed P^0, with respect to all P satisfying the data. Alternatively, by (2.3.28), P *maximizes* the uncertainty $U(P)$ with respect to all P satisfying the data.

The criterion for choosing P is not quite complete, since we have not yet determined the prior probabilities P^0. The prior probabilities represent the state of zero information, i.e. the state corresponding to knowledge of only the set $\mathscr{S}$ of possible outcomes. In the case of the die experiment, P^0 corre-

* By (2.1.47) and (2.1.48), the average value in a long series of throws is highly likely to equal the expectation value on one throw. Thus the expectation value on one throw should be $\langle i \rangle = 4.5$.

sponds to the data that the possible outcomes are $f_1, f_2, \ldots, f_6$. It is fairly clear that the corresponding probabilities should be $p_i^0 = 1/6$ $(i = 1, \ldots, 6)$, since any other assignment would amount to an unwarranted preference for one or another of the f_i.* We can justify this choice of prior probabilities more formally from (2.3.26), which states that the information gained by learning the outcome, when the initial probabilities are p_i^0, is $-\Sigma p_i^0 \ln p_i^0$. Thus, $-\Sigma p_i^0 \ln p_i^0$ represents the missing information or uncertainty in P^0, and hence (by the previous reasoning) the prior probabilities should maximize $-\Sigma p_i^0 \ln p_i^0$ subject to the prior data, i.e. subject only to (1). The result is $p_i^0 = 1/6$.

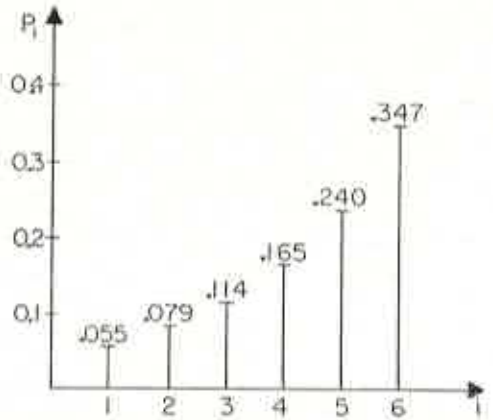

Figure 2.4-1 The probabilities maximizing (3) subject to (1) and (2)

For the die experiment, the probabilities P^0 and P^m satisfy precisely the conditions stated in (2.3.30), so we determine P by maximizing the uncertainty

$$U(P) = -\Sigma p_i \ln p_i \tag{2.4.3}$$

subject to the given data. Extremizing (3) subject to (1) and (2), we obtain (using standard Lagrange multiplier techniques)

$$p_i = A \exp(-\alpha i) \quad (i = 1, \ldots, 6), \tag{2.4.4}$$

where A and the Lagrange multiplier α are chosen so as to satisfy (1) and (2). When this is done, it is found that $\alpha = -0.37$; the result is shown in Figure 1.

* This idea is usually called the "principle of insufficient reason", since $p_i^0 = 1/n$ is chosen due to insufficient reason for preferring one outcome over another. It seems to have been first stated by Laplace.

The above discussion applies to any experiment in which the sample space $\mathscr{S}$ is finite, with the following qualification: (3) is the correct expression for the uncertainty only if the prior data is such that no outcome is preferable over any other, i.e. only if the prior data is symmetric with respect to the outcomes, since otherwise the prior distribution $p_i^0 = 1/n$ is not appropriate. With this qualification, the probability assignment P corresponding to the given data is determined by maximizing (3) subject to the data. If $\mathscr{S}$ is infinite but countable, we may consider the uncertainty $U(P)$ to be the limit, as $n \to \infty$, of the uncertainty in the distribution $(p_1, \ldots, p_n)$. We thus arrive at the following principle of inductive reasoning:

Jaynes' Principle for countable sample spaces. Let $\mathscr{S} = \{\zeta_1, \zeta_2, \ldots\}$ be the set of possible outcomes in some experiment, and assume that the prior description of the experiment is symmetric with respect to the ζ_i. If data D is then given concerning the experiment, the probability assignment $P = (p_1, p_2, \ldots)$ which represents D must maximize the uncertainty (3) with respect to all P satisfying D.

Letting $P(D)$ be the maximizing set of probabilities, the corresponding uncertainty $U[P(D)]$ is the uncertainty in the data D, and $I[P(D); P^0]$ is the information in the data D. Thus, Jaynes' principle extends the information concept from the idea of the information content in a probability distribution to the more direct idea of the information content in a body of data.

Unfortunately, new complications arise when the set $\mathscr{S}$ of possible outcomes is not countable. The previous argument indicates that what we want is to either

(a) minimize $I(P; P^0)$ (equation (2.3.14)) subject to the given data, or
(b) maximize $U(P; P^m; P^0)$ (equation (2.3.28)) subject to the data.

Under either alternative, it is necessary to find some criterion for choosing the prior probabilities P^0. Laplace's principle of insufficient reason would seem to indicate that, if $\mathscr{S} = \{a \leqslant x \leqslant b\}$, then P^0 is given by the density $\varrho^0(x) = \text{const.} = (b-a)^{-1}$. But if we use a new variable $y = g(x)$, then the transformed distribution $\bar{\varrho}^0(y)$ will in general *not* be constant, and hence $\varrho^0(x)$ violates Laplace's principle when the prior data is described in terms of y. Thus, Laplace's principle cannot determine the prior distribution unless we have some criterion for deciding which variable to apply it to. Any variable x to which Laplace's principle applies (i.e. any variable x such that $\varrho^0(x) = \text{const.}$) will be called a **natural variable** for the experiment.

We will not, at this point, give further criteria for the determination of $\varrho^0(x)$. Such criteria have been given by Jaynes,[10] and will be presented in Chapter 3 for the case of classical mechanical systems.

By (2.3.14), if x is a natural variable, then

$$I(P; P^0) = \int_a^b \varrho \ln \varrho \, dx + \ln (b-a).$$

Thus, if the range of x is infinite, and if $\int \varrho \ln \varrho \, dx$ is finite, then $I(P; P^0)$ is infinite. Since in classical statistical mechanics the variable x always has an infinite range, $I(P; P^0)$ is always infinite in statistical mechanics. Thus, alternative (a) is not useful.

So we turn to alternative (b). By (2.3.31), if x is a natural variable and if the state of maximum knowledge is given by a density of the form,

$$\varrho^m(x) = \begin{cases} L^{-1} & (x_1 \leqslant x \leqslant x_1 + L) \\ 0 & (x < x_1 \quad \text{or} \quad x > x_1 + L) \end{cases} \tag{2.4.5}$$

then the uncertainty is

$$U(P) = -\int \varrho(x) \ln [L\varrho(x)] \, dx = -\int \varrho(x) \ln \varrho(x) \, dx - \ln L. \tag{2.4.6}$$

This expression exists whenever $\int \varrho \ln \varrho \, dx$ exists, and hence does not necessarily diverge when the range of x becomes infinite. However, (6) *does* diverge as $L \to 0$. This means that alternative (b) is not useful if we let the state of maximum knowledge be a single point x_0. The easy way out of this dilemma is to assume that L is some small but positive number. Fortunately, in the case of classical statistical mechanics, there are physical reasons (based on the idea that classical mechanics is an approximation to the more "exact" theory of quantum mechanics) for assuming the existence of just such a small number L. In the case of quantum statistical mechanics this question does not arise, since the possible outcomes then form a discrete set.

We thus have the generalization of Jaynes' principle for non-countable sample spaces:

Jaynes' Principle for non-countable sample spaces. Assume that x is a natural variable for the experiment under consideration (i.e. assume that the prior distribution, in terms of x, is $\varrho^0(x) = \text{const.}$), and let the state of maximum knowledge have the form (5). If data D is given concerning the experiment, the probability assignment $\varrho(x)$ which represents D must maximize (6) subject to D.

We have already given an example of Jaynes' principle for countable sample spaces (the die experiment, with the data $\langle i \rangle = 4.5$). As examples of Jaynes' principle for non-countable sample spaces, we will give three useful results. In each case, we will assume that x is a natural variable, and that x ranges over the real axis. The results are established with the help of standard procedures for maximizing integrals subject to given constraints.

If the data is of the form $x \in \mathscr{E}$, where $\mathscr{E}$ is any Lebesgue measurable (i.e. "reasonable") set of finite Lebesgue measure (i.e. finite length) $|\mathscr{E}|$, then Jaynes' principle implies

$$\varrho(x) = \begin{cases} |\mathscr{E}|^{-1} & (x \in \mathscr{E}) \\ 0 & (x \notin \mathscr{E}). \end{cases} \tag{2.4.7}$$

The corresponding uncertainty is

$$U[\varrho(x)] = \ln (|\mathscr{E}|/L). \tag{2.4.8}$$

If the data consists of expectation values of functions $g_j(x)$ $(j = 1, \ldots, k)$,

$$\langle g_j(x) \rangle = G_j \; (j = 1, \ldots, k) \tag{2.4.9}$$

where the G_j are given, then Jaynes' principle implies

$$\varrho(x) = Z^{-1} \exp [-\Sigma \alpha_j g_j(x)], \tag{2.4.10}$$

where the constant Z is determined from $\int \varrho(x) \, dx = 1$, and the α_j are determined from the data (9). The uncertainty in the data (9) is

$$U[\varrho(x)] = \Sigma \alpha_j G_j - \ln (L/Z). \tag{2.4.11}$$

If the given data consists of the mean m and variance σ^2, then Jaynes' principle implies that $\varrho(x)$ is the Gaussian distribution (2.3.32); the corresponding uncertainty is (2.3.33).

REFERENCES

1. A. Papoulis, *Probability, Random Variables and Stochastic Processes* (McGraw-Hill Book Company, New York, 1965).
2. M. J. Lighthill, *Introduction to Fourier Analysis and Generalized Functions* (Cambridge University Press, London, 1958).
3. J. Venn, *The Logic of Chance* (Macmillan, London, 1888) (reprinted, Chelsea, New York, 1963).
4. R. Von Mises, *Probability, Statistics and Truth* (second revised English edition, Macmillan, New York, 1957) (first German edition, 1928).
5. H. Reichenbach, *The Theory of Probability* (University of California Press, Berkeley, 1949) (first German edition, 1935).

6. J.M.Keynes, *A Treatise on Probability* (Macmillan, London, 1921) (reprinted, Harper Torchbooks, New York, 1962).
7. H.Jeffreys, *Scientific Inference* (second edition, Cambridge University Press, Cambridge, 1957).
8. H.Jeffreys, *Theory of Probability* (third edition, Oxford University Press, Oxford, 1961).
9. R.T.Cox, *The Algebra of Probable Inference* (The Johns Hopkins Press, Baltimore, 1961).
10. E.T.Jaynes, IEEE Trans. on Systems Science and Cybernetics, *SSC-4*, 227 (1968).
11. H.E.Kyburg and H.E.Smokler, *Studies in Subjective Probability* (John Wiley and Sons, New York, 1963).
12. C.E.Shannon, *Bell System Tech. J.* **27**, 379 and 623 (1948). Reprinted in C.E.Shannon and W.Weaver, *The Mathematical Theory of Communication* (University of Illinois Press, Urbana, Illinois, 1949).
13. G.Raisbeck, *Information Theory* (The Massachusetts Institute of Technology, 1963).
14. F.M.Reza, *An Introduction to Information Theory* (McGraw-Hill Book Company, New York, 1961).
15. M.S.Pinsker, *Information and Information Stability of Random Variables and Processes*, translated and edited by A.Feinstein (Holden-Day, Inc., San Francisco, 1964).
16. S.Kullback, *Information Theory and Statistics* (Wiley, New York, 1959).
17. E.T.Jaynes, *Phys. Rev.* **106**, 620 (1957); *Phys. Rev.* **108**, 171 (1957).
18. I.J.Good, *Ann. Math. Statist.* **34**, 911 (1963).
19. E.T.Jaynes, in *Statistical Physics* (1962 Brandeis Lectures), ed. by K.W.Ford (W.A. Benjamin, Inc., New York, 1963).

CHAPTER 3

Classical Statistical Mechanics

3.1 THE PROBABILITY DENSITY

ACCORDING TO Chapter 1, classical statistical mechanics is the theory of mechanical systems for which quantum effects are negligible, and the information is less than complete. A complete description at any instant is given by the phase point $x = (q_1, \ldots, q_f, p_1, \ldots, p_f)$ and is called a **state**. An incomplete description is called a **statistical description** or **mixed state**.

If we know the precise initial state and Hamiltonian, then we can make deductive predictions about the system at any time. But if either the precise state or the Hamiltonian is not known, then we can make only inductive predictions, i.e. we are forced to *guess*. According to Section 2.2, probability theory is the appropriate formalism for inductive inferences. Thus, probability theory is the mathematical language for statistical mechanics.

In order to apply probability theory to incompletely specified mechanical systems, we must find an appropriate sample space $\mathscr{S}$, field of events $\mathbf{F}$, and probability assignment P (see Section 2.1). The obvious choice of $\mathscr{S}$ is phase space. Thus an outcome, denoted in Section 2.1 by ζ, means a phase point $x \in \mathscr{S}$. The simplest and most useful choice of $\mathbf{F}$ is the collection of all Lebesgue measurable (i.e. "reasonable") sets of phase points, so that *event* means any Lebesgue measurable set of phase points.*

A probability assignment P means the assignment of a probability $P(\mathscr{E})$ to every $\mathscr{E} \in \mathbf{F}$, such that postulates (2.1.1), (2.1.2), (2.1.3) are satisfied. Since mechanics is a dynamical theory, P will generally be different at different times. We will denote the entire probability assignment at the instant t by $P(\ ;t)$; the probability of $\mathscr{E}$ at time t will then be denoted by $P(\mathscr{E};t)$.

In classical statistical mechanics, $P(\ ;t)$ is usually a continuous probability assignment, possessing a Lebesgue integrable **probability density** (also

* As explained in Section 2.1 (see the footnote on page 18), it is necessary to restrict $\mathbf{F}$ to only the Lebesgue measurable sets. Readers unfamiliar with the notions of "Lebesgue measurable" and "Lebesgue measure" may replace these terms by "reasonable" and "phase volume" without undue loss of precision.

called a **probability distribution**) $\varrho(x, t)$ such that for any $\mathscr{E} \in \mathbf{F}$

$$P(\mathscr{E}; t) = \int_{x \in \mathscr{E}} \varrho(x, t)\, dx.^* \tag{3.1.1}$$

We may think of the phase point as a $2f$-dimensional random variable, since x is indeed a set of $2f$ numbers whose values are determined by the outcome x (see Section 2.1). Thus, x plays a dual role: on the one hand, any particular x is an outcome, and on the other hand x is a random variable.

Any numerical function of the outcome x is an r.v. For instance, the Hamiltonian $H(x)$, the total momentum, and the first coordinate q_1 are one-dimensional r.v. The coordinates $(q_1, \ldots, q_f)$ form an f-dimensional r.v.

Some authors define a new concept called the *ensemble;* this concept is roughly equivalent to the idea of the probability distribution $\varrho(x, t)$. The ensemble is the cause of much needless confusion in statistical mechanics, and is best a superfluous idea, so we will not use it.

In some cases, $P(\ ; t)$ may be discrete, with probabilities $p_1, p_2, \ldots$ assigned to a discrete set of points $x_1, x_2, \ldots$ In this case, there is no Lebesgue measurable function $\varrho(x, t)$ satisfying (1). However, we may still prescribe $P(\ ; t)$ by (1) provided we understand $\varrho(x, t)$ to be the generalized function

$$\varrho(x, t) = \Sigma\, p_i\, \delta(x - x_i(t)). \tag{3.1.2}$$

Exact mechanics corresponds to information of the form "the phase point at time t is $x_i(t)$, with probability 1", so that mechanics is a special case of statistical mechanics, with probability density $\varrho(x, t) = \delta(x - x_i(t))$.

According to Sections 2.2 and 2.4, we determine the probability distribution from the given information. In statistical mechanics, this is accomplished by using Jaynes' principle (Section 2.4) and the dynamics of the mechanical system. The dynamical conditions will be discussed in Section 3.3; Jaynes' principle will be applied in Section 3.6.

3.2 A SIMPLE EXAMPLE

Simple examples are always helpful in physics, but especially so in statistical mechanics since the ideas of phase space and probability densities in phase space are fairly abstract.

* Strictly speaking, all integrals in this book are Lebesgue integrals. However, the reader may think of the integrals as Riemann integrals without undue loss of precision, since for Riemann integrable functions, the Lebesgue integral equals the Riemann integral.

The simplest system which illustrates most of the ideas in this chapter is a bead of mass m moving freely along a length L of wire, bouncing elastically from each end. We will call this example the "particle in a box".

The mechanical state of the particle in a box is given by $x = (q, p)$, where q is the position and p is the momentum. Phase space $\mathscr{S}$ is the strip $0 \leqslant q \leqslant L$, $-\infty < p < \infty$. The Hamiltonian is $H(x) = p^2/2m$ $(x \in \mathscr{S})$.

A simple probability density at some instant $t = 0$ is

$$\varrho(q, p, 0) = \begin{cases} 2/bL & (0 \leqslant q \leqslant L/2 \quad \text{and} \quad 0 \leqslant p \leqslant b) \\ 0 & (\text{otherwise}). \end{cases} \tag{3.2.1}$$

This density is illustrated in Figure 1 (one of the advantages of the particle in a box is that we can draw phase space on a sheet of paper). The density (1) is non-negative and normalized, as it should be.

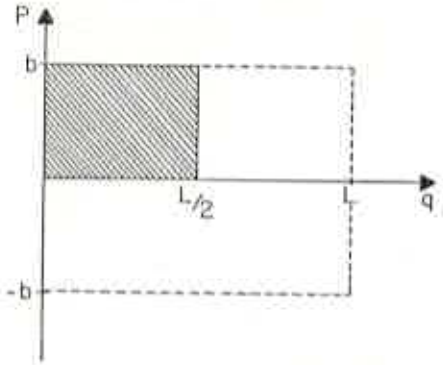

Figure 3.2-1 The probability density (1). Inside the shaded region, $\varrho = 2/bL$. Outside, $\varrho = 0$

It is essential to understand that ϱ does not represent the mechanical state of the system; ϱ represents only the observer's *knowledge* (or data) about the state of the system. The distribution (1) does *not* say that the phase point of the system is somehow smeared out over the rectangle of Figure 1. The actual mechanical state of the system is given by some precise point inside this rectangle. The only thing that is smeared out is the observer's *knowledge:* the observer does not know just where the phase point is, so he represents what he *does* know by ϱ.

According to (2.4.7), the density (1) corresponds to the data that "the particle is in the left half of the box at $t = 0$, with momentum between 0 and b". We will elaborate on this point in Section 3.6.

Putting (1) into (3.1.1), we see that

$$P(\{q_1 < q < q_2\}; 0) = 2(q_2 - q_1)/L \quad (q_2 < L/2), \tag{3.2.2}$$

$$P(\{p_1 < p < p_2\}; 0) = (p_2 - p_1)/b \quad (0 < p_1 < p_2 < b), \tag{3.2.3}$$

$$P(\{q_1 < q < q_2\} \cap \{p_1 < p < p_2\}; 0) = 2(q_2 - q_1)(p_2 - p_1)/bL \quad (q_2 < L/2 \text{ and } 0 < p_1 < p_2 < b), \tag{3.2.4}$$

$$P(\{H(x) < E\}; 0) = \sqrt{2mE}/b \quad (E < b^2/2m), \tag{3.2.5}$$

$$P(\{E < H(x) < E + \Delta E\}; 0) \simeq \sqrt{m/2b^2E}\,\Delta E \quad (\Delta E \text{ small, and } E + \Delta E < b^2/2m). \tag{3.2.6}$$

By (2), (3), and (4), if $q_2 < L/2$ and $0 < p_1 < p_2 < b$ then

$$P(\{q_1 < q < q_2\} \cap \{p_1 < p < p_2\}; 0) = P(\{q_1 < q < q_2\}; 0)\, P(\{p_1 < p < p_2\}; 0). \tag{3.2.7}$$

We can easily show that (7) holds for any q_1, q_2, p_1, p_2. Thus the r.v. q and p are statistically independent (compare (2.1.36)). We could have seen this intuitively from Figure 1, by the following reasoning: information about q does not affect predictions about p (for instance, if we are given $q = L/4$, we can still only say that $0 \leqslant p \leqslant b$); but according to the inductive inference interpretation of probability, statistical independence of q and p *means* that new information about q does not affect predictions about p; thus, q and p are statistically independent.

According to (6) and (2.1.19), the probability density for the r.v. $H(x)$ (i.e. for the energy E) is

$$\varrho(E, t = 0) = \begin{cases} \sqrt{m/2b^2E} & (E \leqslant E_{\max}) \\ 0 & (E > E_{\max}), \end{cases} \tag{3.2.8}$$

where $E_{\max} \equiv b^2/2m$. Thus, within the range $(0 \leqslant E \leqslant E_{\max})$, smaller energies are more probable than larger energies. For instance, using (5),

$$P(\{E \leqslant E_{\max}/2\}; 0) = 1/\sqrt{2} \simeq 0.7. \tag{3.2.9}$$

3.3 DYNAMICS OF THE PROBABILITY ASSIGNMENT

Information about a mechanical system at time t_0 is obviously relevant to predictions at any other time t_1. Thus, the dynamics of the mechanical system relates the probability assignment $P(\ ; t_0)$ in some manner to $P(\ ; t_1)$,

i.e. the dynamics translates the data from t_0 to t_1. The object of this Section is to investigate this time-translation of the data. We will assume that the Hamiltonian is time-independent and precisely known.

We first derive the basic condition which the dynamics imposes on the time-dependence of $P(\ ; t)$. Corresponding to any initial phase point x_0 at time t_0, let $X(t|x_0, t_0)$ represent the phase point at time t; thus, x_0 evolves into $x = X(t|x_0, t_0)$ during t_0 to t. The set of points $x = X(t|x_0, t_0)$, for

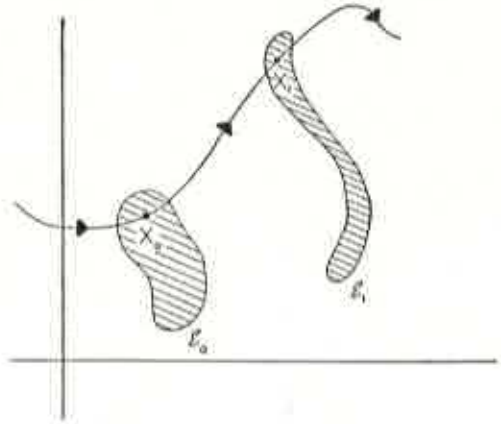

Figure 3.3-1 Phase space diagram of two sets $\mathscr{E}_0$ and $\mathscr{E}_1$, such that $\mathscr{E}_0$ evolves into $\mathscr{E}_1$ during t_0 to t_1. A portion of one phase path is shown, with phase point x_0 at t_0 and $x_1 \equiv (t_1|x_0, t_0$ at t_1

fixed (x_0, t_0), and for $-\infty < t < \infty$, is called a **phase path**; thus the moving phase point traces out a phase path. The mechanical motion defines a one-to-one transformation of phase space: for fixed t_0 and t, there is exactly one phase point $x = X(t|x_0, t_0)$ corresponding to each $x_0 \in \mathscr{S}$, and conversely for each $x \in \mathscr{S}$ (and for fixed t_0 and t) there is exactly one x_0 such that $x = X(t|x_0, t_0)$.

Let $\mathscr{E}_0$ be any Lebesgue measurable set of phase points, and define $\mathscr{E}_1$ as the set of all points $X(t_1|x_0, t_0)$ such that $x_0 \in \mathscr{E}_0$, for fixed t_0 and t_1. That is, $\mathscr{E}_0$ evolves into $\mathscr{E}_1$ during t_0 to t_1 (see Figure 1). From the definition of $\mathscr{E}_0$ and $\mathscr{E}_1$, *if* $\mathscr{E}_0$ occurred at t_0 *then* $\mathscr{E}_1$ must occur at t_1. In symbols,

$$P(\mathscr{E}_1; t_1 \mid \mathscr{E}_0; t_0) = 1,$$

from which (using (2.1.9))

$$P(\mathscr{E}_1; t_1 \cap \mathscr{E}_0; t_0) = P(\mathscr{E}_0; t_0). \tag{3.3.1}$$

The definition of $\mathscr{E}_0$ and $\mathscr{E}_1$, along with the fact that $X(t|x_0, t_0)$ defines a one-to-one transformation of $\mathscr{S}$, implies that *if* $\mathscr{E}_1$ occurs at t_1, *then* $\mathscr{E}_0$ must have occurred at t_0. Thus

$$P(\mathscr{E}_0; t_0 \mid \mathscr{E}_1; t_1) = 1,$$

from which (again using (2.1.9))

$$P(\mathscr{E}_1; t_1 \cap \mathscr{E}_0; t_0) = P(\mathscr{E}_1; t_1). \tag{3.3.2}$$

Finally, (1) and (2) imply $P(\mathscr{E}_1; t_1) = P(\mathscr{E}_0; t_0)$. Thus the basic condition which the dynamics imposes is:

> If the Lebesgue measurable set of phase points $\mathscr{E}_0$ evolves into $\mathscr{E}_1$ during t_0 to t_1, then $P(\mathscr{E}_1; t_1) = P(\mathscr{E}_0; t_0)$. (3.3.3)

We next find the condition which (3) imposes on the evolution of $\varrho(x, t)$. Denote the Lebesgue measure of any measurable set $\mathscr{E}$ by $|\mathscr{E}|$; that is,

$$|\mathscr{E}| \equiv \int_{x \in \mathscr{E}} dx. \tag{3.3.4}$$

A theorem of classical mechanics[1] states that, if $\mathscr{E}_0$ evolves into $\mathscr{E}_1$ during t_0 to t_1, and if the variables $x = (q_1, \ldots, p_f)$ are canonical (i.e. such that Hamilton's equations hold), then

$$|\mathscr{E}_0| = |\mathscr{E}_1|. \tag{3.3.5}$$

Outline of the proof of (3.3.5): Define

$$|\mathscr{E}| \equiv \int_{x \in \mathscr{E}} dq_1, \ldots, dq_f, dp_1, \ldots, dp_f.$$

Under an arbitrary change of variables $(q_1, \ldots, p_f) \to (q_1', \ldots, p_f')$, we have

$$|\mathscr{E}| = \int_{x' \in \mathscr{E}'} J(q_1', \ldots, p_f')\, dq_1', \ldots, dp_f'$$

where

$$J(q_1', \ldots, p_f') \equiv \frac{\partial(q_1, \ldots, p_f)}{\partial(q_1', \ldots, p_f')}$$

is the Jacobian of the transformation, and where $\mathscr{E}'$ is the region into which $\mathscr{E}$ transforms under the transformation. It can be shown[1] that if $(q_1, \ldots, p_f) \to (q_1', \ldots, p_f')$ is a *canonical* transformation, then $J(q_1', \ldots, p_f') = 1$. Hence,

for canonical transformations,

$$|\mathscr{E}| = \int_{x' \in \mathscr{E}'} dq_1', \ldots, dp_f' \equiv |\mathscr{E}'|.$$

It can be shown[1] that, if the variables $x = (q_1, \ldots, p_f)$ are canonical, then the transformation $x \to x' \equiv X(t_1|x, t_0)$ is canonical, for any fixed t_0 and t_1. That is, the natural motion of any Hamiltonian system amounts to a continuous sequence of canonical transformations. Thus, for any t_0 and t_1,

$$|\mathscr{E}_0| \equiv \int_{x \in \mathscr{E}_0} dx = \int_{x' \in \mathscr{E}_1} dx' \equiv |\mathscr{E}_1|,$$

where $\mathscr{E}_1$ is the region into which $\mathscr{E}_0$ transforms under the transformation $x \to x' \equiv X(t_1|x, t_0)$.

Now assume that $\mathscr{E}_0$ is a small region containing the point x_0. Then, by (3.1.1),

$$P(\mathscr{E}_0; t_0) = \int_{x \in \mathscr{E}_0} \varrho(x, t_0)\, dx \simeq \varrho(x_0, t_0)\, |\mathscr{E}_0|, \tag{3.3.6}$$

where the approximation is good if $\mathscr{E}_0$ is sufficiently small (more precisely, if the maximum value of $|x - x_0|$ is sufficiently small, for all $x \in \mathscr{E}_0$). If $\mathscr{E}_0$ evolves into $\mathscr{E}_1$ during t_0 to t_1, then the point $x_1 \equiv X(t_1|x_0, t_0)$ must be in $\mathscr{E}_1$. Thus

$$P(\mathscr{E}_1; t_1) = \int_{x \in \mathscr{E}_1} \varrho(x, t_1)\, dx \simeq \varrho(x_1, t_1)\, |\mathscr{E}_1|. \tag{3.3.7}$$

Condition (3) and equations (5), (6) and (7) imply $\varrho(x_1, t_1) \simeq \varrho(x_0, t_0)$. We can make the approximations as accurate as desired simply by choosing $|\mathscr{E}_0|$ sufficiently small. Thus

$$\varrho(x_1, t_1) = \varrho(x_0, t_0), \quad \text{where} \quad x_1 \equiv X(t_1 \mid x_0, t_0). \tag{3.3.8}$$

This is the basic dynamical condition on ϱ; it says that if x_0 evolves into x_1 during t_0 to t_1, and if the variables $x = (q_1, \ldots, p_f)$ are canonical, then ϱ has the same value at (x_1, t_1) as it has at (x_0, t_0).

An **integral of the motion** means a phase function $g(x, t)$ such that, for all x_0, t_0, and t_1,

$$g(x_1, t_1) = g(x_0, t_0) \quad \text{where} \quad x_1 \equiv X(t_1 \mid x_0, t_0). \tag{3.3.9}$$

That is, for any fixed initial point x_0 at time t_0, the value $g[X(t \mid x_0, t_0), t]$ of the function g does not change as $X(t|x_0, t_0)$ moves along the phase path.

A **constant of the motion** means a time-independent integral of the motion, i.e. a function $g(x)$ such that, for all x_0, t_0 and t_1,

$$g(x_1) = g(x_0) \quad \text{where} \quad x_1 \equiv X(t_1 \mid x_0, t_0). \tag{3.3.10}$$

For example, the Hamiltonian $H(x)$ is a constant of the motion. If no external forces act on the system, then the total momentum vector and the angular momentum vector are constants of the motion, and $\mathbf{R} - \mathbf{P}t/M$ ($\mathbf{R}$ = center of mass, $\mathbf{P}$ = total momentum, M = total mass) is an integral of the motion. An interesting question in the general theory of mechanical systems (and a question which is important for statistical mechanics) is: how many constants of the motion exist which are both Lebesgue integrable and functionally independent, and what is the nature of these constants of the motion? This question is easily answered for 2-particle systems. But for the more interesting case of N interacting particles, the answer is largely unknown. A recent result by Sinai[2] suggests that, for interacting systems confined by external forces, the Hamiltonian is the *only* constant of the motion which is Lebesgue integrable and functionally independent of all other constants of the motion. Sinai has proven this result for a system of N hard spheres bouncing elastically from the walls of a box, for any $N \geqslant 2$. It is not known whether this result also holds for more general interactions.

It is known from mechanics[1] that if $g(x, t)$ is an integral of the motion, and if all the appropriate derivatives exist, then

$$\frac{\partial g}{\partial t} + [g, H] = 0, \tag{3.3.11}$$

where the square brackets mean the Poisson bracket:

$$[u(x), v(x)] \equiv \sum_{i=1}^{f} \left(\frac{\partial u}{\partial q_i} \frac{\partial v}{\partial p_i} - \frac{\partial v}{\partial q_i} \frac{\partial u}{\partial p_i} \right). \tag{3.3.12}$$

Equation (8) says that $\varrho(x, t)$ is an integral of the motion. Thus, by (11),

$$\frac{\partial \varrho}{\partial t} + [\varrho, H] = 0. \tag{3.3.13}$$

Equation (13) is called **Liouville's equation**; it is the basic differential equation for $\varrho(x, t)$.

Most dynamical problems in statistical mechanics are equivalent to finding (perhaps only approximately or partially) $\varrho(x, t)$ in terms of $\varrho(x, 0)$. One way of doing this is to use (8) directly: replacing (x_1, t_1) by $(x_0, 0)$, and

replacing (x_0, t_0) by (x, t), we get

$$\varrho(x_0, 0) = \varrho(x, t) \quad \text{where} \quad x_0 \equiv X(0 \mid x, t). \tag{3.3.14}$$

But, for systems having time-independent Hamiltonians, $X(t_1 \mid x_0, t_0)$ depends on t_0 and t_1 only through the difference $t_1 - t_0$. Thus $X(0 \mid x, t) = X(-t \mid x, 0)$, and (14) becomes

$$\varrho(x, t) = \varrho[X(-t \mid x, 0), 0]. \tag{3.3.15}$$

Thus, if $X(-t \mid x, 0)$ is known for all x and t (i.e. if the solution to Hamilton's equations is known!), then $\varrho(x, t)$ may be found in terms of $\varrho(x, 0)$ from (15).

Another way to find $\varrho(x, t)$ in terms of $\varrho(x, 0)$ is by solving Liouville's equation (13). The general method of solving linear first-order partial differential equations such as (13) is discussed in any book on partial differential equations (see, e.g., Ref. 3). In addition, various expansion techniques, perturbation techniques, diagram techniques, etc., have been developed for the specific purpose of solving, or partially or approximately solving, Liouville's equation (see, e.g., Refs. 4–7). Since this is a book about concepts rather than techniques, we will not discuss these methods.

We will mention one further result concerning the solution of Liouville's equation:

> Let $\phi_1(x, t), \ldots, \phi_k(x, t)$ be a set of differentiable integrals of the motion, and let $f(u_1, \ldots, u_k)$ be any differentiable function of k variables. Then $f[\phi_1(x, t), \ldots, \phi_k(x, t)]$ satisfies Liouville's equation. (3.3.16)

The proof will be left to the reader. Property (16) simply says that any function of integrals of the motion is itself an integral of the motion. According to (16), if we can express $\varrho(x, 0)$ in terms of known integrals of the motion:

$$\varrho(x, 0) = f[\phi_1(x, 0), \ldots, \phi_k(x, 0)],$$

then we can immediately write $\varrho(x, t)$ as

$$\varrho(x, t) = f[\phi_1(x, t), \ldots, \phi_k(x, t)]. \tag{3.3.17}$$

Equation (17) says that the probability distribution of the integrals of the motion is preserved in time. Thus, it is easy to translate in time any given information about the integrals of the motion.

The particle in a box illustrates most of these ideas. Let $\varrho(x, 0)$ be given by (3.2.1). For this example, it is easier to use (8) than to solve Liouville's

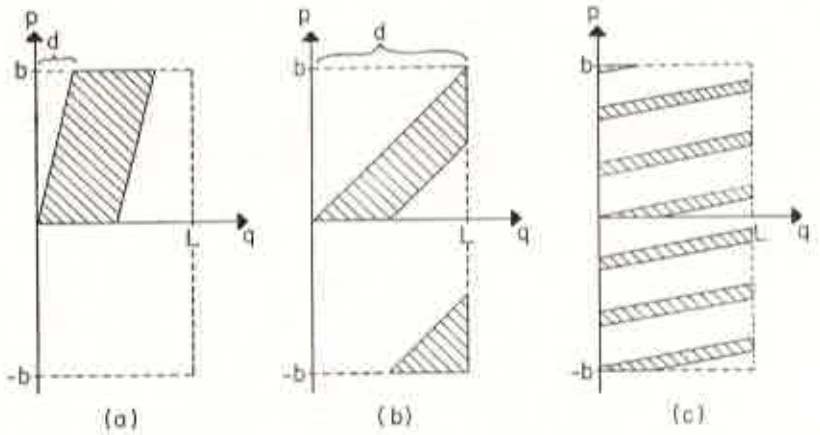

Figure 3.3-2 The probability distribution for the particle in a box, at three different times. Inside the shaded region, $\varrho = 2/bL$. Outside, $\varrho = 0$. The distance $d = bt/m$

equation. Using (8), we see that after a short time t the distribution $\varrho(x, t)$ is as shown in Figure 2a, with $d \equiv bt/m$. After a longer time, the density begins to "reflect" from the wall $q = L$, as shown in Figure 2b (drawn for the instant at which $d = bt/m = L$). At much longer times (such that $bt/m \gg L$), the distribution has the shape shown in Figure 2c.

As time progresses, the original rectangle of Figure 3.2-1 distorts into a finer and finer filament which stretches back and forth across phase space. This phenomenon, first noticed by Gibbs and known as "Gibbs phase mixing", will be discussed in detail in Chapter 5. Despite the fact that the shape of the distribution in Figure 2c is quite different from the original rectangular shape, by (5) the area of the shaded region in Figure 2c equals the area $bL/2$ of the original rectangle. Despite the periodicity of the mechanical system (the particle is just bouncing back and forth between 0 and L), the probability assignment never repeats. We see that *statistical descriptions of mechanical systems may have features which are qualitatively different from mechanical descriptions.*

Let $t = mL/4b$, so that $d = L/4$ (see Figure 2a); at this time,

$$\begin{aligned} P(\{q < L/2\}; t) &= 0.75 \\ P(\{p > 0\}; t) &= 1. \end{aligned} \tag{3.3.18}$$

Thus the particle is certainly moving toward the right, but there is only a 75% probability that it is still in the left-hand half of the box. Let $t = mL/b$, so

that $d = L$ (see Figure 2b). At this time,

$$\begin{aligned} P(\{q < L/2\}; t) &= 0.25 \\ P(\{p > 0\}; t) &= 0.75. \end{aligned} \tag{3.3.19}$$

It seems fairly clear from Figure 2c that, for $t \gg mL/b$,

$$\begin{aligned} P(\{q < L/2\}; t) &\simeq 0.5 \\ P(\{p > 0\}; t) &\simeq 0.5, \end{aligned} \tag{3.3.20}$$

where the approximation gets better as t gets larger.

Property (3) implies that the probability distribution for the energy is time-independent (simply choose $\mathcal{E}_0$ to be the region between two "energy surfaces" $H(x) = E$ and $H(x) = E + \Delta E$).

The dynamics of $P(\ ; t)$ described in this Section serve only to *translate* data from one time t_0 to another time t_1. If, on the other hand, we obtain new data at some instant $\bar{t}$ between t_0 and t_1, and if we wish to base our predictions for t_1 on *all* the available data, then the assignment $P(\ ; t_1)$ arising from $P(\ ; t_0)$ and (3) is not appropriate. We should instead base our predictions on $\bar{P}(\ ; t_1)$ arising from $\bar{P}(\ ; \bar{t})$, where $\bar{P}(\ ; \bar{t})$ is the probability assignment at $\bar{t}$ which reflects not only the *old* data (i.e. the data described at $\bar{t}$ by $P(\ ; \bar{t})$) but also the new data obtained at $\bar{t}$. If the given data at $\bar{t}$ is really new, i.e. not predictable from $P(\ ; \bar{t})$, then $\bar{P}(\ ; \bar{t})$ will be different from $P(\ ; \bar{t})$; that is, the probability assignment undergoes a sudden jump or collapse when we obtain new data. Thus $\varrho(x, t)$ obeys Liouville's equation at all times except those instants when new data is obtained; at such instants, ϱ undergoes a sudden collapse.

3.4 EXPECTATION VALUES, REDUCED DISTRIBUTIONS, CORRELATIONS

Any phase function $g(x)$ is a random variable, since the numerical value of $g(x)$ is determined by the outcome x. The expectation value (see (2.1.40)) of $g(x)$ over the probability assignment $P(\ ; t)$ will be denoted $\langle g(x) \rangle_t$:

$$\langle g(x) \rangle_t \equiv \int g(x)\, \varrho(x, t)\, dx.^* \tag{3.4.1}$$

As discussed in Section 3.1, each component of x is an r.v., and x itself is an r.v. Thus $\varrho(x, t)$ is the joint distribution of the $2f$ r.v. $q_1, \ldots, p_f$, and reduced distributions may be defined as in Section 2.1 (see (2.1.31)).

---

* All integrals are over the entire range of the variables, unless explicitly stated otherwise.

Assuming that the system consists of N point particles,* the phase point is $x = (\mathbf{q}_1, \ldots, \mathbf{q}_N, \mathbf{p}_1, \ldots, \mathbf{p}_N)$, where $\mathbf{q}_i$ and $\mathbf{p}_i$ are now *vectors* representing the position and momentum of the ith particle.** The reduced distribution for $(\mathbf{q}_i, \mathbf{p}_i)$ at time t is

$$\varrho_i^{(1)}(\mathbf{q}_i, \mathbf{p}_i, t) = \int \varrho(x, t)\, d\mathbf{q}_1, \ldots, d\mathbf{q}_{i-1}\, d\mathbf{q}_{i+1}, \ldots, d\mathbf{q}_N\, d\mathbf{p}_1, \ldots, d\mathbf{p}_{i-1}\, d\mathbf{p}_{i+1}, \ldots, d\mathbf{p}_N. \tag{3.4.2}$$

We may generalize equation (2) in an obvious way to obtain an expression for the reduced distribution $\varrho_{ij\ldots k}^{(r)}(\mathbf{q}_i, \mathbf{q}_j, \ldots, \mathbf{q}_k, \mathbf{p}_i, \mathbf{p}_j, \ldots, \mathbf{p}_k)$ for the r particles numbered $i, j, \ldots, k$. Using only $\varrho_i^{(1)}$, we can answer all questions about the ith particle alone. For instance, from $\varrho_i^{(1)}$ we can calculate the probability that $\mathbf{q}_i$ is in some 3-dimensional region R. On the other hand, we must use $\varrho_{ij}^{(2)}$ to answer such "2-body questions" as the probability that $|\mathbf{q}_i - \mathbf{q}_j| > d$, or the probability that $\mathbf{q}_i \in R_1$ given $\mathbf{q}_j \in R_2$.

Often, $\varrho(\mathbf{q}_1, \ldots, \mathbf{q}_N, \mathbf{p}_1, \ldots, \mathbf{p}_N, t)$ is symmetric under any particle exchange $(\mathbf{q}_i, \mathbf{p}_i) \leftrightarrow (\mathbf{q}_j, \mathbf{p}_j)$. By Liouville's equation, $\varrho(x, t)$ is symmetric whenever both $\varrho(x, 0)$ and $H(x)$ are symmetric. But we usually determine $\varrho(x, 0)$ from the initial data (see Section 3.6); thus $\varrho(x, t)$ is symmetric whenever the Hamiltonian and the initial data are symmetric under particle exchange.

If $\varrho(x, t)$ is symmetric, then by (2) the function $\varrho_i^{(1)}(\mathbf{q}, \mathbf{p}, t)$ is independent of i. In this case, $\varrho_i^{(1)}$ will be written $\varrho^{(1)}$ and called the "1-body probability distribution". Similarly, $\varrho_{ij\ldots k}^{(r)}$ may be written $\varrho^{(r)}$ and called the "r-body probability distribution". Whenever we use the notation $\varrho^{(r)}$, it is to be understood that ϱ is symmetric.

With $\varrho^{(1)}$, we can answer any question about any single particle. For example, for any 1-body phase function $\phi(\mathbf{q}_i, \mathbf{p}_i)$ we have

$$\langle \phi(\mathbf{q}_i, \mathbf{p}_i) \rangle_t = \int d\mathbf{q}\, d\mathbf{p}\, \phi(\mathbf{q}, \mathbf{p})\, \varrho^{(1)}(\mathbf{q}, \mathbf{p}, t). \tag{3.4.3}$$

Using (3) and the linearity of the expectation value operation,

$$\left\langle \sum_{i=1}^{N} \phi(\mathbf{q}_i, \mathbf{p}_i) \right\rangle_t = N \int d\mathbf{q}\, d\mathbf{p}\, \phi(\mathbf{q}, \mathbf{p})\, \varrho^{(1)}(\mathbf{q}, \mathbf{p}, t) \tag{3.4.4}$$

* This restriction is made only for convenience. The entire presentation may be easily generalized to the case of an arbitrary classical system, possessing f degrees of freedom.

** Thus the number of degrees of freedom is $f = 3N$.

so that from $\varrho^{(1)}$ alone we can find the expectation value of any phase function of the form

$$\Phi(x) = \sum_{i=1}^{N} \phi(\mathbf{q}_i, \mathbf{p}_i). \tag{3.4.5}$$

Any phase function of the form (5) is called a **sum function**.

Similarly, we can answer questions about any pair of particles by means of $\varrho^{(2)}$. Furthermore, from $\varrho^{(2)}$ we can find the expectation value of any phase function of the form

$$\Phi(x) = \sum_i \sum_j \phi(\mathbf{q}_i, \mathbf{q}_j, \mathbf{p}_i, \mathbf{p}_j). \tag{3.4.6}$$

Since most phase functions of interest in statistical mechanics have the form (5) or (6), and since (see Section 3.5) we are usually interested only in expectation values, we do not generally need to know the complete probability density $\varrho(x, t)$; the reduced distributions $\varrho^{(1)}$ and $\varrho^{(2)}$ are sufficient.

There is an interesting relation between $\varrho^{(1)}$ and the expectation value of the number density. By the **number density** we mean a function $n(\mathbf{q})$ such that the *exact* number of particles in any region R of 3-dimensional coordinate space is

$$\text{number of particles in } R = \int_{\mathbf{q} \in R} n(\mathbf{q})\, d\mathbf{q}. \tag{3.4.7}$$

The function $n(\mathbf{q})$ is obviously a different function for different phase points x; thus we will denote it by $n(\mathbf{q}; x)$. Hence the number density is a mechanical phase function, dependent on $\mathbf{q}$ as a parameter. The only such function satisfying (7) is

$$n(\mathbf{q}; x) = n(\mathbf{q}; \mathbf{q}_1, \ldots, \mathbf{q}_N) = \sum_{i=1}^{N} \delta(\mathbf{q} - \mathbf{q}_i).* \tag{3.4.8}$$

By (4) and (8), the expectation value of the particle density, at any fixed point $\mathbf{q}$, is

$$\langle n(\mathbf{q}; x) \rangle_t = N \int d\mathbf{p}\, \varrho^{(1)}(\mathbf{q}, \mathbf{p}, t) \equiv N \varrho_{\mathbf{q}}^{(1)}(\mathbf{q}, t) \tag{3.4.9}$$

where $\varrho_{\mathbf{q}}^{(1)}$ is the reduced probability distribution for the position of one particle alone.

Now, (9) says that the *expected number density* is proportional to the 1-body *probability density*. The similarity of the terms "number density" and "probability density", plus the close relation (9) between these two quantities makes for lots of confusion between these two radically different concepts.

* Thus $n(\mathbf{q}; x)$ is not really a function at all—it is a generalized function.

The confusion is enhanced by the fact (see Section 3.5) that the *exact* number of particles

$$n(R; x) \equiv \int_{\mathbf{q}\in R} d\mathbf{q}\, n(\mathbf{q}; x) \tag{3.4.10}$$

in any sufficiently large region R is highly likely to be approximately equal to the *expected* number of particles in R:

$$\langle n(R; x)\rangle_x = N\int_{\mathbf{q}\in R} \varrho_q^{(1)}(\mathbf{q}, t)\, d\mathbf{q} \simeq N\varrho_a^{(1)}(\mathbf{q}, t)\,|R|, \tag{3.4.11}$$

where

$$|R| \equiv \int_R d\mathbf{q}.$$

That is, $n(R; x)$ is quite likely to be approximately equal to $N|R|\,\varrho_q^{(1)}(\mathbf{q}, t)$. (This is true if N is large and R is not too small; see Section 3.5). Despite this probable proportionality between $\varrho_q^{(1)}(\mathbf{q}, t)$ and the integrated value (10) of $n(\mathbf{q}; x)$, the two concepts are radically different; $n(R, x)$ is a mechanical phase function (or, in probabilistic terminology, a random variable), whereas $\varrho_q^{(1)}(\mathbf{q}, t)$ is a reduced probability distribution or, by (9), an expectation value. We may see the distinction even more clearly from the fact that $n(\mathbf{q}; x)$ and $n(R; x)$ both have a meaning within exact mechanics, while $\varrho_q^{(1)}(\mathbf{q}, t)$ has a meaning only within a *probabilistic description* of mechanical systems. The conceptual difficulty raised here is typical of difficulties which arise in statistical mechanics and kinetic theory (particularly in discussions of the Boltzmann transport equation), and for which the reader should be on guard.

From (2.1.37), each particle is **statistically independent** of every other particle at time t if and only if

$$\varrho(x, t) = \prod_{i=1}^{N} \varrho_i^{(1)}(\mathbf{q}_i, \mathbf{p}_i, t). \tag{3.4.12}$$

As in Section 2.1, we say particles i and j are **correlated** if they are not statistically independent. According to the inductive inference interpretation of probability (Section 2.2), particles i and j are statistically independent if and only if any new information about particle i is irrelevant to predictions about particle j.

The superficially similar ideas of dynamical independence and statistical independence are logically distinct: one is a mechanical idea, while the other is a statistical idea. A set of particles is dynamically independent if they don't

exert forces on each other, i.e. if

$$H(x) = \Sigma H_i(\mathbf{q}_i, \mathbf{p}_i). \tag{3.4.13}$$

They are statistically independent if new data about one doesn't change predictions about the others. For example, given two strongly interacting particles in a box, and given the data that $\mathbf{q}_1 \in R_1$ and $\mathbf{q}_2 \in R_2$, the appropriate probability distribution for the positions $\mathbf{q}_1$ and $\mathbf{q}_2$ is (see Section 3.6)

$$\varrho_a^{(2)}(\mathbf{q}_1, \mathbf{q}_2) = \begin{cases} \dfrac{1}{|R_1|\,|R_2|} & (\mathbf{q}_1 \in R_1 \quad \text{and} \quad \mathbf{q}_2 \in R_2) \\ 0 & (\text{otherwise}). \end{cases}$$

Thus, $\mathbf{q}_1$ and $\mathbf{q}_2$ are statistically independent even though the particles are strongly interacting. Again, let the system consist of N free, non-interacting particles, and let the given information be that the total energy is between E_1 and E_2:

$$E_1 \leqslant \Sigma\, \mathbf{p}_i^2/2m \leqslant E_2.$$

Then the particles are correlated; for instance, the new data $\mathbf{p}_1^2/2m = E_2$ immediately implies $\mathbf{p}_i^2/2m = 0$ $(i \neq 1)$, so that new data about one particle is relevant to predictions about the others. Hence, this system is dynamically independent but correlated.

From (12), (13), and Liouville's equation, we can show that, if the particles are statistically independent at $t = 0$, and if the particles are dynamically independent, then they are statistically independent for all t. Thus, if $\varrho(x, t)$ contains inter-particle correlations then either $\varrho(x, 0)$ must have contained correlations or the Hamiltonian contains interactions. That is, correlations at time t arise either from correlations at $t = 0$ or from interactions.

We will say that a system of particles is **weakly correlated** if (12) is approximately true, i.e. if

$$\varrho(x, t) = \Pi\, \varrho_i^{(1)}(\mathbf{q}_i, \mathbf{p}_i, t) + \mathbf{c}(x, t) \tag{3.4.14}$$

where the correlation function $c(x, t)$ defined by (14) satisfies $\int |c(x, t)|\, dx \ll 1$. We will say that a system is **weakly interacting** if (13) is approximately true, i.e. if

$$H(x) = \Sigma H_i(\mathbf{q}_i, \mathbf{p}_i) + H'(x) \tag{3.4.15}$$

where the interaction term $H'(x)$ is in some sense small, or unimportant, compared to ΣH_i (we will not try to be terribly precise about this). We can show*

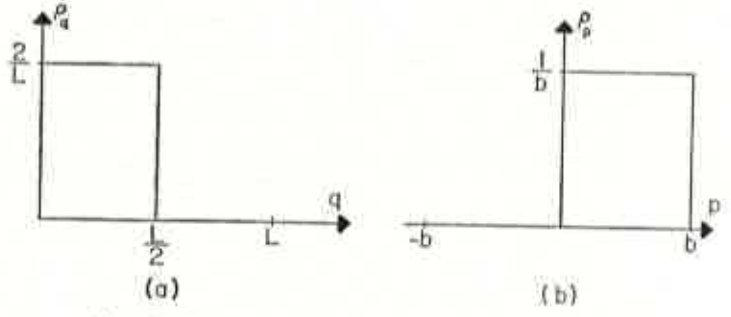

Figure 3.4-1 Reduced distributions for q and p at $t = 0$

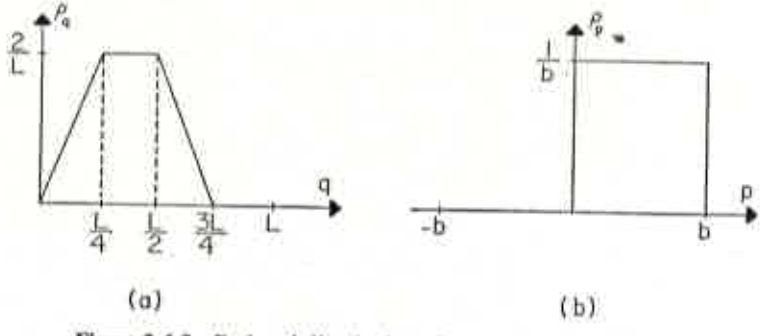

Figure 3.4-2 Reduced distributions for q and p at $t_1 = mL/4b$

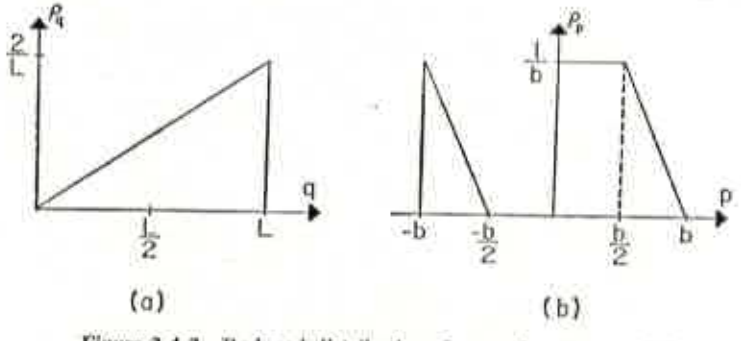

Figure 3.4-3 Reduced distributions for q and p at $t_2 = mL/b$

that, if the system is weakly interacting and if it is weakly correlated at $t = 0$, then it is weakly correlated at any time t such that $|t|$ is not too large.

We can illustrate most of the ideas of this Section with the particle in a box. From Figure 3.2-1,

$$\langle q\rangle_{t=0} = L/4, \quad \langle p\rangle_{t=0} = b/2,$$

and the initial reduced distributions are as shown in Figure 1. The r.v. q and p are statistically independent at $t = 0$.

Furthermore, at the times $t_1 = mL/4b$ and $t_2 = mL/b$ (see Figures 3.3-2a and 3.3-2b),

$$\langle q\rangle_{t_1} = 3L/8, \quad \langle p\rangle_{t_1} = b/2,$$

$$\langle q\rangle_{t_2} = 2L/3, \quad \langle p\rangle_{t_2} = b/12.$$

Figures 2 and 3 show the reduced distributions at t_1 and t_2. From Figure 3.3-2 q and p are correlated at t_1 and t_2. For instance, the new information that "$p < 0$ at t_2" implies (using Figure 3.3-2b) that $q > L/2$ at t_2; that is, new information about p is relevant to predictions made about q, so p and q are correlated.

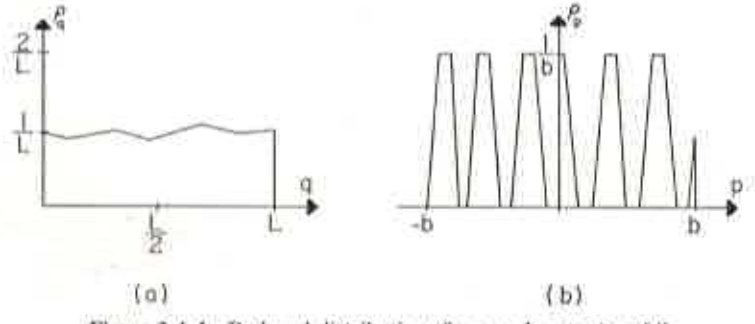

Figure 3.4-4 Reduced distributions for q and p at $t \gg mL/b$

According to Figure 3.3-2c, at long times ($t \gg mL/b$)

$$\langle q\rangle_t \simeq L/2, \quad \langle p\rangle_t \simeq 0,$$

and the reduced distributions have the forms shown in Figure 4. Thus, the expectation values "relax" to constants for $t \gg mL/b$, as shown in Figure 5.

---

* The easiest way is to write $\varrho(0) = \Pi\varrho_i^{(1)}(0) + \lambda c(0)$ and $H = \Sigma H_i + \lambda H'$, and use (3.3.15) to show that $\varrho(t) = \Pi\varrho_i^{(1)}(t) +$ terms of order λ.

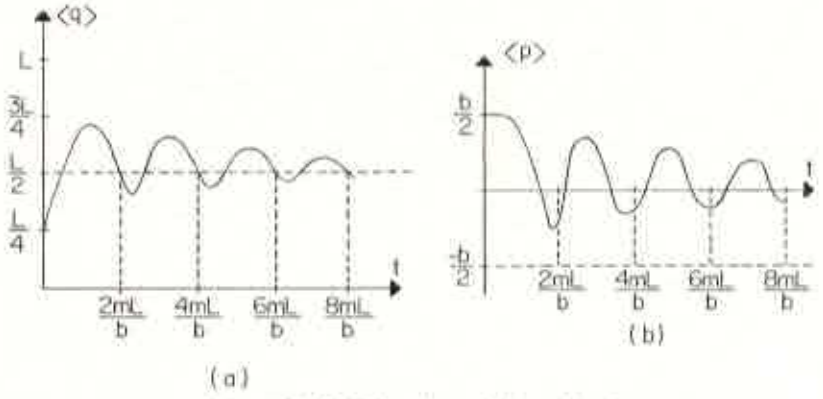

Figure 3.4-5 Relaxation of $\langle q \rangle$ and $\langle p \rangle$

Note that, whereas the probability distribution $\varrho(q, p, t)$ never relaxes (ϱ just continues to stretch into a finer and finer filament, so that $\lim_{t \to \infty} \varrho(q, p, t)$ does not exist), the expectation values $\langle q \rangle_t$ and $\langle p \rangle_t$ do relax to constants. This behavior is typical of statistical mechanics, and will be discussed in detail in Chapter 5. Once again, we see that mechanical and statistical descriptions may be qualitatively different: the underlying (but unknown) mechanical phase point undergoes periodic motion, while $\langle q \rangle$ and $\langle p \rangle$ relax to constants.

3.5 THE EFFECT OF LARGE N

It is practically impossible to treat systems of more than a few particles by exact mechanics. In fact, a famous and nearly intractable problem of exact mechanics is the "3-body problem": three particles interacting through gravitational forces. Even if we could carry out the exact mechanical calculations for a many-body system, it is nearly impossible to experimentally obtain the initial data (i.e. the precise initial phase point) required for such calculations to yield precise predictions. Thus one effect of large N is that exact mechanics can no longer be applied, and it becomes necessary to use some other scheme, such as statistical mechanics.

We will show in this Section that many quantities of interest in statistical mechanics are highly predictable when N is large. This is the most significant effect of large N.

According to the Tchebycheff inequality (2.1.45), if an r.v. X has a small dispersion σ, then we can predict (with a high degree of confidence) that X lies within a small range of $m \equiv \langle X \rangle$. As discussed in Section 2.1 (p. 25), a better measure of predictability is the relative dispersion σ/m.*

Many phase functions of interest in many-body systems are sum functions, of the form (3.4.5). We will show that, if

$$\varrho(x, t) = \prod_{i=1}^{N} \varrho^{(1)}(\mathbf{q}_i, \mathbf{p}_i, t) \tag{3.5.1}$$

(i.e. if ϱ is symmetric and uncorrelated), then the relative dispersion of the sum function $\Phi(x) = \Sigma \phi(\mathbf{q}_i, \mathbf{p}_i)$ is

$$\frac{\sigma}{m} = \frac{\sigma_1}{m_1 \sqrt{N}} \tag{3.5.2}$$

where m_1 and σ_1^2 are the mean and variance of $\phi(\mathbf{q}_1, \mathbf{p}_1)$ at time t. *Proof:* Each function $\phi(\mathbf{q}_i, \mathbf{p}_i)$ is an r.v. Under the stated assumptions, these r.v. are statistically independent with common mean and variance. Their average is $\Phi(x)/N$, and, by (2.1.48), the variance of the average is σ_1^2/N. Thus the variance of $\Phi(x)$ is $\sigma^2 = N\sigma_1^2$. By (3.4.4), the mean of $\Phi(x)$ is $m = Nm_1$. Thus σ/m is given by (2).

Equation (2) says that the relative dispersion of any sum function may be made as small as desired simply by choosing N sufficiently large. Thus, provided ϱ has the form (1), any sum function is highly predictable if N is large enough.

In the more general case that $\varrho(x, t)$ is symmetric and weakly correlated (see (3.4.14)), sum functions are again highly predictable since (2) is still approximately valid. Recall from Section 3.4 that $\varrho(x, t)$ is weakly correlated if $\varrho(x, 0)$ is weakly correlated and if $H(x)$ is weakly interacting. Now, for many systems of physical interest there exists a set of canonical coordinates and momenta $(\mathbf{w}_1, \ldots, \mathbf{w}_N, \mathbf{J}_1, \ldots, \mathbf{J}_N)$ such that $H(\mathbf{w}_1, \ldots, \mathbf{J}_N)$ is weakly interacting. It often turns out that the initial distribution $\varrho(\mathbf{w}_1, \ldots, \mathbf{J}_N, 0)$, expressed in terms of these weakly interacting variables, is weakly correlated (for the reason, see (3.6.7)). Thus, in most cases of physical interest, $\varrho(\mathbf{w}_1, \ldots, \mathbf{J}_N, t)$ is weakly correlated for all t, and hence predictions about 1-body sum functions $\Sigma \phi(\mathbf{w}_i, \mathbf{J}_i)$ are highly predictable. It is fairly clear that we could make a similar argument about "2-body sum functions" of the form $\Sigma\Sigma \phi(\mathbf{w}_i, \mathbf{w}_j, \mathbf{J}_i, \mathbf{J}_j)$, or about "$n$-body sum functions" with $n \ll N$. But the only quantities which are measurable in many-body systems are n-body sum

* If X can take on positive or negative values, it may happen that $m = 0$. In this case the relative dispersion is not terribly useful.

functions with $n \ll N$, since other phase functions are usually too complicated to be measured. Thus, we conclude that *in most many-body systems of physical interest, all measurable quantities are highly predictable.*

Examples of sum functions are:

$$\text{total momentum} = \Sigma\, \mathbf{p}_i,$$

and, assuming the particles are of equal mass,

$$\text{center of mass} = N^{-1} \Sigma\, \mathbf{q}_i,$$

$$\text{kinetic energy} = \Sigma\, \mathbf{p}_i^2/2m.$$

If the system is non-interacting and has a symmetric Hamiltonian, then the total energy is a sum function; if the particles are pairwise interacting and if $H(x)$ is symmetric, then the total energy is a "2-body sum function" of the form (3.4.6).

Another example of a sum function is the integrated number density. Using (3.4.8) and (3.4.10), the exact number of particles in the 3-dimensional region R, when the phase point is x, is

$$n(R; x) = \int_R d\mathbf{q} \sum_{i=1}^{N} \delta(\mathbf{q} - \mathbf{q}_i) = \sum_{i=1}^{N} \int_R d\mathbf{q}\, \delta(\mathbf{q} - \mathbf{q}_i) = \sum_{i=1}^{N} \theta(\mathbf{q}_i). \tag{3.5.3}$$

Thus $n(R; x)$ is a sum function, with $\theta(\mathbf{q}_i)$ being the step function

$$\theta(\mathbf{q}_i) = \begin{cases} 1 & (\mathbf{q}_i \in R) \\ 0 & (\mathbf{q}_i \notin R). \end{cases} \tag{3.5.4}$$

We can show that, if $\varrho(x, t)$ is symmetric under particle exchange,

$$\langle \theta(\mathbf{q}_i) \rangle = \langle \theta^2(\mathbf{q}_i) \rangle = \int_R \varrho_q^{(1)}(\mathbf{q}, t)\, d\mathbf{q};$$

consequently the relative dispersion of $\theta(\mathbf{q}_i)$ is

$$\frac{\sigma_1}{m_1} = \left[\frac{1 - P(\{\mathbf{q}_i \in R\}; t)}{P(\{\mathbf{q}_i \in R\}; t)} \right]^{1/2} \tag{3.5.5}$$

where

$$P(\{\mathbf{q}_i \in R\}; t) = \int_R \varrho_q^{(1)}(\mathbf{q}, t)\, d\mathbf{q}. \tag{3.5.6}$$

Equations (2), (3) and (5) imply that, if N is large and if $P(\{\mathbf{q}_i \in R\}; t)$ is not too small (specifically, we must have $P(\{\mathbf{q}_i \in R\}; t) \gg N^{-1}$), then the relative dispersion of $n(R; x)$ is small. In other words, if N is large and R is not too small, then the actual number of particles in R is quite likely to be approximately equal to the expected number of particles in R.

As pointed out in Section 3.4, the close relationship between $n(R; x)$ and its expectation value $N \int_R \varrho_q^{(1)}(\mathbf{q}, t)\, d\mathbf{q}$ leads to a certain amount of confusion. It is easy, for example, to believe (mistakenly) that $N \int_R \varrho_q^{(1)}(\mathbf{q}, t)\, d\mathbf{q}$ is a *mechanical phase function*, whose value is approximately the number of particles in R. Instead, $N \int_R \varrho_q^{(1)}(\mathbf{q}, t)\, d\mathbf{q}$ is an *expectation value*, whose numerical value is highly likely (provided N is large and R is not too small) to be approximately equal to the number of particles in R. Thus, it is possible (but not very probable) that, on any given trial, the exact value of $n(R; x)$ will differ by a large amount from its mean $N \int_R \varrho_q^{(1)}(\mathbf{q}, t)\, d\mathbf{q}$. That is, an experimental measurement of the number of particles in R might yield a number which deviates widely from the expected value $N \int_R \varrho_q^{(1)}(\mathbf{q}, t)\, d\mathbf{q}$.

3.6 THE INITIAL DISTRIBUTION

The purpose of statistical mechanics is to make predictions about mechanical systems, based on incomplete data. Usually, we want to predict what will happen at some instant t_1, when the data is collected at some other instant $t = 0$. We have seen, in Section 3.3, how to translate a probability assignment from $t = 0$ to t_1; in this Section, we find a method for determining a probability assignment from the given initial data. The basic scheme of classical statistical mechanics will then be complete, since we will have a way of translating data collected at $t = 0$ into predictions at any other time t_1.

Since classical phase space is a continuum, and since the data is incomplete, the obvious way to determine $P(\quad ; t = 0)$ is through Jaynes' principle for non-countable sample spaces (Section 2.4). To apply Jaynes' principle, we must first find the distribution $\varrho^0(x)$ which represents complete ignorance. We will show that, if the variables x are canonical, then $\varrho^0(x) = \text{const.}$; i.e. the canonical variables are the natural variables (see Section 2.4) for any mechanical system.

In order to determine $\varrho^0(x)$, we will use an idea which has been stated in the following way by Jaynes:[9] "If we merely specify complete initial ignorance, we cannot hope to obtain any definite prior distribution, because such a statement is too vague to define any mathematically well-posed problem. We are defining what we mean by complete ignorance far more precisely if we can specify a set of operations which we recognize as transforming the problem into an equivalent one ..."

Suppose that we are given a closed mechanical system with Hamiltonian $H(x)$, and told only that "5 seconds ago, at $t = -5$, the exact phase point was in the phase space $\mathscr{S}$". That is we have complete ignorance about the phase point at $t = -5$. What can we say about the system at $t = 0$? The answer is that we cannot say anything about it, except that the phase point is still in $\mathscr{S}$. That is, we have complete ignorance at $t = 0$. (If this were not true, then we could gain information about physical systems simply by waiting a long time, without ever making any measurements!) Thus, $\varrho^0(x)$ is a state such that, if $\varrho^0(x)$ is used as an initial distribution at t_0 (in the above example, $t_0 = -5$) and then translated (by the methods of Section 3.3) into a distribution at t_1 (in the above example, $t_1 = 0$), the distribution at t_1 is again the state of complete ignorance $\varrho^0(x)$. But, assuming that $\mathscr{S}$ is described by canonical variables x, Liouville's equation (3.3.13) describes the time-evolution of any probability density. Thus, invariance of $\varrho^0(x)$ under time-translation implies

$$[\varrho^0(x), H(x)] = 0. \tag{3.6.1}$$

We will now generalize the above argument. Suppose that we are given a system which is closed from $t = 0$ to ∞, with Hamiltonian $H(x)$ during this time, and that we are given the following data: "At $t = -5$, the exact phase point was in $\mathscr{S}$, and during $t = -5$ to $t = 0$ the Hamiltonian was $H'(x)$." (For example, the system during $t = 0$ to ∞ might be an electron in a box in zero field, and the data might be "at $t = -5$, the electron was in the box, and during $t = -5$ to $t = 0$ an electric field of 5 volt/meter was directed downward inside the box".) What can we say about the system at $t = 0$? The answer, once again, is that we can't say anything about it except that the phase point is still in $\mathscr{S}$. Thus, when we translate the distribution $\varrho^0(x)$ from t_0 to t_1 under the influence of an arbitrary Hamiltonian $H'(x)$, the state of knowledge at t_1 is still just $\varrho^0(x)$. Assuming that the variables x are canonical, this implies

$$[\varrho^0(x), H'(x)] = 0 \tag{3.6.2}$$

for every possible Hamiltonian $H'(x)$ on $\mathscr{S}$.

In the general theory of mechanical systems, any differentiable phase function $f(x)$ is a possible Hamiltonian. Thus (2) says that $[\varrho^0(x), f(x)] = 0$ for every differentiable $f(x)$. But the only differentiable distribution $\varrho^0(x)$ satisfying this condition is

$$\varrho^0(x) = \text{const.} = |\mathscr{S}|^{-1}$$

where $|\mathscr{S}| \equiv \int dx$ is the total Lebesgue measure of $\mathscr{S}$ (we assume, for the moment, that $|\mathscr{S}|$ is finite). Thus, at least for finite $|\mathscr{S}|$, the canonical variables are the natural variables.

In case $|\mathscr{S}|$ is infinite (as it always is in classical mechanics, since there is no a priori bound on the momenta), it is reasonable to take $\varrho(x, 0)$ as the limit, as $|\mathscr{S}| \to \infty$, of that distribution corresponding to the given data on a finite phase space. Thus the natural variables are the canonical variables even if $|\mathscr{S}|$ is infinite.

We will now determine the distribution corresponding to maximum permissible knowledge. In doing this, we will find it necessary to use Heisenberg's uncertainty principle from quantum mechanics. Consider a system having a single degree of freedom. An experimental measurement of q and p always leads to a result of the form "q lies in the range $q_1 < q < q_2$, and p lies in $p_1 < p < p_2$," where the "spreads" $\Delta q \equiv q_2 - q_1$ and $\Delta p \equiv p_2 - p_1$ are non-zero. (Note that the argument thus far does not rely on quantum mechanics.) Classically, Δq and Δp may be any positive numbers, no matter how small. The idea of the uncertainty principle is that there *is* a limit to the smallness of Δq and Δp, or rather to the smallness of their product. The uncertainty principle says that, no matter what experimental apparatus is used, $\Delta q\, \Delta p \geqslant h$ where h is Planck's constant. If the system has f degrees of freedom, then the experimental "spreads" $\Delta q_i, \Delta p_i$ must satisfy $\Delta q_i\, \Delta p_i \geqslant h$ $(i = 1, \ldots, f)$. Thus an experimental measurement can at best determine that $x \in \mathscr{E}$, where

$$|\mathscr{E}| = \prod_{i=1}^{f} \Delta q_i\, \Delta p_i = h^f. \tag{3.6.3}$$

Thus, the state of maximum knowledge has the form

$$\varrho^m(x) = \begin{cases} |\mathscr{E}|^{-1} & (x \in \mathscr{E}) \\ 0 & (x \notin \mathscr{E}) \end{cases}$$

where $|\mathscr{E}| = h^f$.

We have now shown that, in classical statistical mechanics, the uncertainty $U[\varrho(x)]$ in the distribution $\varrho(x)$ is given by (2.4.6), where x represents the canonical variables and where $L = h^f$.

It appears that if the uncertainty principle did not exist, it would have to be invented in order to make logically consistent use of inductive reasoning in classical statistical mechanics! It is true that Jaynes' principle is independent of the *numerical value* of the parameter L in (2.4.6), since (2.4.6) is maximized whenever $-\int \varrho \ln \varrho \, dx$ is. But Jaynes' principle does depend on the *existence* of some lower bound L to the phase volume $|\mathscr{E}|$ within which x can be experimentally located. Without such a lower bound the uncertainty $U(P)$ is infinite for every continuous probability assignment P, and hence we cannot find P by maximizing $U(P)$.

The above argument leading to the minimum volume (3) tacitly assumes that the particles are microscopically distinguishable. But suppose that the system consists of N microscopically indistinguishable particles. That is, taking $N = 2$, suppose that the state $x = (\mathbf{q}, \mathbf{q}', \mathbf{p}, \mathbf{p}')$ cannot possibly (by any means available to man) be distinguished from the state $x' = (\mathbf{q}', \mathbf{q}, \mathbf{p}', \mathbf{p})$. Many naturally occurring systems consist of indistinguishable particles. For example, a collection of electrons, or photons, or alpha particles, is of this type.

If the system consists of N indistinguishable particles, then the smallest phase volume $|\mathscr{E}|$ within which we can experimentally locate x is not $|\mathscr{E}| = h^f$, but is instead $|\mathscr{E}| = N! \cdot h^f$.* For example, taking $N = 2$, the most precise measurement possible can only yield data of the form "the two particles are in a region $\mathscr{E}_1$ around $(\mathbf{q}, \mathbf{q}', \mathbf{p}, \mathbf{p}')$, or else they are in a region $\mathscr{E}_2$ around $(\mathbf{q}', \mathbf{q}, \mathbf{p}', \mathbf{p})$, where $|\mathscr{E}_1| = |\mathscr{E}_2| = h^6$." Thus we can experimentally determine only that the particles are within $\mathscr{E}_1 \cup \mathscr{E}_2$, where $|\mathscr{E}_1 \cup \mathscr{E}_2| = 2h^6$ (assuming $\mathscr{E}_1 \cap \mathscr{E}_2 = 0$).** Thus, the minimum phase volume h^f allowed by the uncertainty principle must be multiplied by the number $N!$ of indistinguishable permutations among the N particles.

In the case of r distinguishable species of particles, with N_k indistinguishable particles of the kth species ($k = 1, \ldots, r$), the number of indistinguishable permutations is $N_1! \ldots N_r!$ (for instance, if the system consists of N_1 electrons and N_2 protons, then the number of indistinguishable permutations is $N_1!\, N_2!$). Summarizing, the principle for choosing $P(\quad ; t = 0)$ is:

> **Jaynes' Principle** for classical statistical mechanics. If initial data D is given relative to a system of r distinguishable species with N_k indistinguishable particles of each species,

* Note that $f = 3N$.

** We will not worry too much about the possibility that $\mathscr{E}_1 \cap \mathscr{E}_2 \neq 0$, since the set of points $(\mathbf{q}, \mathbf{q}', \mathbf{p}, \mathbf{p}')$ on which $\mathscr{E}_1$ and $\mathscr{E}_2$ overlap is usually of very small probability.

> and if the variables x are canonical, then the initial probability assignment is determined by maximizing the uncertainty
>
> $$U[\varrho(x, 0)] = -\int \varrho(x, 0) \ln [h^{3N} N_1! \cdots N_r!\, \varrho(x, 0)]\, dx$$
>
> subject to D, where $N = N_1 + \cdots + N_r$. (3.6.4)

The basic formalism of classical statistical mechanics is now complete, since we now have mathematical rules for translating any given data at $t = 0$ into predictions at t. The two basic principles are (3.6.4), which determines $P(\quad ; t = 0)$, and (3.3.3), which determines $P(\quad ; t)$ in terms of $P(\quad ; t = 0)$.

As an example, suppose that we possess the following three measuring instruments for the particle in a box: M_1 is a light bulb which is on only when the particle is in the left half of the box; M_2 is a light bulb which comes on when the particle hits the left wall and turns off when the particle hits the right wall (thus M_2 is on only when $p > 0$); M_3 is an impact device at one of the walls, which measures the magnitude of the momentum with a coarseness b (that is, M_3 indicates which energy shell $nb < |p| < (n + 1)b$ the particle is in). Suppose that, at $t = 0$, M_1 and M_2 are on and M_3 registers $0 < |p| < b$. Applying Jaynes' principle, and recalling (2.4.7), the initial distribution is just (3.2.1). Predictions about future values shown by M_1, M_2, and M_3 are then made using the ideas of Section 3.3.

We will present several interesting consequences of Jaynes' principle.

> Let the data be $E_1 \leqslant H(x) \leqslant E_2$, where E_1 and E_2 are given. Then
>
> $$\varrho(x, t = 0) = \begin{cases} |\mathscr{E}|^{-1} & (x \in \mathscr{E}) \\ 0 & (x \notin \mathscr{E}) \end{cases}$$
>
> where $\mathscr{E}$ is the region defined by the data. (3.6.5)

This result follows from (2.4.7). The region $\mathscr{E}$ is called an energy shell, since it lies between the two energy surfaces defined by $H(x) = E_1$ and $H(x) = E_2$. The distribution (5) is called the **microcanonical distribution.** Thus, if we are given the information that the energy lies in some specified range E_1 to E_2, we should base our predictions on the microcanonical distribution.

If the data is $\langle H(x)\rangle = E$, where E is given, then

$$\varrho(x, t=0) = Z^{-1} \exp[-\beta H(x)]$$

where

$$Z \equiv \int \exp[-\beta H(x)]\, dx \equiv Z(\beta),$$

and β is chosen to satisfy the data. The defining condition for β may be expressed as

$$E = -\frac{\partial}{\partial \beta} \ln Z(\beta). \tag{3.6.6}$$

This result is just a special case of (2.4.10). The defining equation for β may be verified by carrying out the derivative on the right hand side. The distribution (6) is Gibbs' famous **canonical distribution**, and $Z(\beta)$ is the **partition function**. Thus, if we are given the expectation value of the energy, we should base our predictions on the canonical distribution.

Let the data be symmetric with respect to the different particles, and assume that the data is entirely expressible as conditions on the 1-body probability distribution $\varrho^{(1)}(\mathbf{q}, \mathbf{p}, t=0)$. The initial distribution then has the form

$$\varrho(x, 0) = \prod_{i=1}^{N} \varrho^{(1)}(\mathbf{q}_i, \mathbf{p}_i, t=0). \tag{3.6.7}$$

For the proof, see Appendix B. Thus uncorrelated initial distributions arise whenever the data can be expressed in terms of $\varrho^{(1)}$ alone. For example, if the data consists of the expectation values of sum functions (see (3.4.5)), then the initial distribution is uncorrelated.

Let the data consist of information about the differentiable and functionally independent* phase functions $\phi_1(x), \ldots, \phi_k(x)$. Then $\varrho(x, 0)$ is functionally dependent on the $\phi_i(x)$:

$$\varrho(x, 0) = f[\phi_1(x), \ldots, \phi_k(x)]. \tag{3.6.8}$$

For the proof, see Appendix B. The number of functions $\phi_i(x)$ cannot be greater than $2f$, since the functions could not then be functionally independent in the $2f$-dimensional space $\mathscr{S}$. As an example of (8), the data of (5) involves only the energy and hence the corresponding distribution depends on x only through $H(x)$; the same is true of (6).

* Two functions are functionally independent if the numerical value of one of them does not determine a unique numerical value for the other. More precisely, two functions $\phi_1(q_1, \ldots, p_f)$ and $\phi_2(q_1, \ldots, p_f)$ are functionally *dependent* if there exists a functional relation $F(\phi_1, \phi_2) = 0$ for all $(q_1, \ldots, p_f)$. For example, $\phi_1 = q_1$ and $\phi_2 = (q_1)^2$ are functionally dependent, but $\phi_1 = q_1$ and $\phi_2 = q_2$ are functionally independent.

One interesting and potentially very important application of Jaynes' principle concerns the case in which data is gathered over an extended period of time rather than at a single instant $t = 0$. Suppose that several phase functions $g_1(\mathbf{q}; x), \ldots, g_m(\mathbf{q}; x)$, each possibly dependent on a continuous position parameter $\mathbf{q}$ (see, for instance, (3.4.8)) have been observed (by a macroscopic observer) at all points $\mathbf{q}$ in some 3-dimensional region R, and at all times from $t = -\tau$ to $t = 0$. Assume that N is sufficiently large that the observed data may be identified with expectation values (see Section 3.5). The given data then has the form

$$\langle g_k(\mathbf{q}; x)\rangle_t = G_k(\mathbf{q}, t) \quad (\mathbf{q} \in R,\ -\tau \leqslant t \leqslant 0), \quad (k = 1, 2, \ldots, m), \tag{3.6.9}$$

where the $G_k(\mathbf{q}, t)$ are known throughout R and for all times from $t = -\tau$ to $t = 0$. Using (3.3.15), the expectation value at time t is

$$\langle g_k(\mathbf{q}; x)\rangle_t \equiv \int g_k(\mathbf{q}; x)\, \varrho(x, t)\, dx = \int g_k(\mathbf{q}; x)\, \varrho[X(-t|x), 0]\, dx,$$

where $X(t|x)$ represents the phase point at time t corresponding to the initial point x. Under the change of variables $x \to y = X(-t|x)$, we have $x = X(t|y)$ and (by the theorem (3.3.5)) $dx = dy$, so the above integral becomes

$$\langle g_k(\mathbf{q}; x)\rangle_t = \int g_k[\mathbf{q}; X(t|y)]\, \varrho(y, 0)\, dy.$$

Thus, we can express the expectation value of $g_k(\mathbf{q}; x)$ at time t as the initial expectation value of the function $g_k[\mathbf{q}; X(t|x)]$, and the given data (9) becomes

$$\langle g_k[\mathbf{q}; X(t|x)]\rangle_0 = G_k(\mathbf{q}, t) \quad (\mathbf{q} \in R,\ -\tau \leqslant t \leqslant 0,\ k = 1, \ldots, m). \tag{3.6.10}$$

We are now in a position to apply Jaynes' principle to the initial data (10).

By partitioning the space-time observation region ($\mathbf{q} \in R$, $-\tau \leqslant t \leqslant 0$) into a large number of small space-time "cells", the continuous initial data (10) may be expressed, approximately, in the discrete form (2.4.9). Jaynes' principle then leads to an initial distribution function of the form (2.4.10). In the limit as the size of the cells approaches zero, this approximation becomes exact and the corresponding initial distribution becomes*

$$\varrho(x, 0) = \frac{1}{Z} \exp\left\{\int_R d\mathbf{q} \int_{-\tau}^{0} dt \sum_{k=1}^{m} \lambda_k(\mathbf{q}, t)\, g_k[\mathbf{q}; X(t|x)]\right\}. \tag{3.6.11}$$

Equation (11) is Jaynes' **generalized canonical distribution**; it represents a broad generalization of Gibbs' canonical distribution (6). The generalized

partition functional Z (the normalization factor in (11)) is given by

$$Z[\lambda_1(\mathbf{q},t),\ldots,\lambda_m(\mathbf{q},t)] \equiv \int \exp\left\{\int_R d\mathbf{q}\int_{-\tau}^{0} dt \sum_{k=1}^{m} \lambda_k(\mathbf{q},t) g_k[\mathbf{q}; X(t|x)]\right\} dx \tag{3.6.12}$$

The generalized Lagrange multipliers $\lambda_k(\mathbf{q},t)$ ($\mathbf{q} \in R$, $-\tau \leqslant t \leqslant 0$) are chosen to satisfy the given data (10). It turns out* that the defining condition for the $\lambda_k(\mathbf{q},t)$ may be expressed as

$$G_k(\mathbf{q},t) = \frac{\delta}{\delta\lambda_k(\mathbf{q},t)} \ln Z[\lambda_1(\mathbf{q},t),\ldots,\lambda_m(\mathbf{q},t)], \tag{3.6.13}$$

where $\delta/\delta\lambda_k(\mathbf{q},t)$ represents a functional derivative. Equations (11), (12), and (13) are a natural generalization of Gibbs' result (6); we may apply these equations to any equilibrium or non-equilibrium situation for which the given data has the form (9). Predictions about any quantity $f(x)$ at any time $t > 0$ are then based on $\varrho(x,t)$, found by solving Liouville's equation subject to the initial distribution (11).

3.7 WHY DOES STATISTICAL MECHANICS WORK

Statistical mechanics is the scheme which has been worked out for making the most reasonable predictions about the behavior of mechanical systems, based on incomplete information. But ordinary experience shows that predictions based on incomplete information do not always turn out to be correct. Why, then, does statistical mechanics work?

Suppose we are given initial data D, and that we wish to predict the value of some phase function $g(x)$ at time t. Assume that statistical mechanics leads to the conclusion that (on the basis of D) it is highly likely that $g(x)$ at time t will equal $\langle g\rangle_t$ to within experimental accuracy. That is, letting δ be the experimental error in measuring $g(x)$,

$$P(\{|g(x) - \langle g\rangle_t| < \delta\}; t) \simeq 1. \tag{3.7.1}$$

We will then say that the prediction $|g(x) - \langle g\rangle_t| < \delta$ is **highly plausible**; we have seen in Section 3.5 that most statistical mechanical predictions are of this type. Suppose that we now proceed to actually measure $g(x)$ at time t, obtaining a value g_m.

Suppose that g_m does lie in the predicted range $|g_m - \langle g\rangle_t| < \delta$. This is hardly surprising; it merely means that an inductive inference which was highly plausible turned out to be correct. But *why* was it correct? Well, it was correct because inductive reasoning usually works. To push the questioning one step further by asking "why does inductive reasoning work?" would lead us beyond science and into philosophy. In science, one ordinarily assumes the validity of inductive reasoning, since if inductive reasoning *didn't* work science would not be possible in the first place. Thus, the most complete answer we can give to the question posed in this chapter is that *statistical mechanics works because inductive reasoning works.*

Now suppose that g_m does *not* lie in the predicted range, and that on reruns of the experiment the measured values of $g(x)$ at time t continue to lie outside the predicted range.* (On each trial, the data D must be the same, i.e. the uncertainties in the initial preparation must be the same.) In this case, the prediction is experimentally wrong. Nevertheless, it is still true that statistical mechanics yields the *best* prediction (i.e. the least biased prediction) possible on the basis of the data. In other words, it is not statistical mechanics which is at fault; rather, the data is wrong or inadequate, or the underlying mechanical theory (classical mechanics in this chapter) is inadequate, or else a mathematical mistake was made in the statistical mechanical calculations (for example, an invalid approximation or an inconsistent series expansion may have been used).

When we experimentally test a statistical mechanical prediction, we are not really testing statistical mechanics; statistical mechanics is simply a special case of inductive reasoning, and in science one does not ordinarily test inductive reasoning (although science itself is in a sense a test of inductive reasoning). Thus if the measured value of a highly predictable phase function falls consistently outside the predicted range, and if no mathematical errors have been made, then we are forced to conclude that either the data or the mechanics is inadequate. Statistical mechanics is not invalid in such a case, far from it; in fact, it is just when statistical mechanical reasoning leads to an incorrect prediction that statistical mechanics is the most useful, for we are then in a position to learn something new about the physical world.

As a very simple, non-mechanical example, suppose that the experiment is the simultaneous tossing of 1000 dice, and that we wish to predict the total number of spots showing. Let the initial data be only that each die has 6 sides, with i spots on the ith side, so that we initially have complete ignorance about the outcome. On the basis of this data, the probability distribution

* If in a large number of trials only a few measured values fall outside the predicted range, then the prediction is considered to be correct since the prediction was not certain anyway, but was only highly plausible.

$P(i_1, i_2, \ldots, i_{1000})$ (where i_r represents the number of spots showing on the rth die) must be uncorrelated, since the data contains no information relating the outcome on one die to the outcome on another die. Thus $P(i_1, \ldots, i_{1000}) = \Pi P_1(i_r)$, where $P_1(i)$ is the probability distribution for a single die. The distribution $P_1(i)$ satisfying Jaynes' principle for the given data is just $P_1(i) = 1/6$ $(i = 1, 2, \ldots, 6)$. This distribution has mean and variance $m_1 = 3.5$, $\sigma_1^2 = 2.750$. Thus the total number of spots has mean and variance $m = 1000 m_1 = 3500$, $\sigma^2 = 1000\,\sigma_1^2 = 2750$. The dispersion is $\sigma \simeq 50$. By the Tchebycheff inequality, the probability that the total number of spots will lie outside of the range 3300 to 3700 is less than $1/(4)^2 = 6.25\%$. The probability that the result will lie outside the range 3100 to 3900 is less than $1/(8)^2 \simeq 1.5\%$. Thus, on the basis of the data, it is highly likely (more precisely, the probability is greater than 98.5%) that the total number of spots will lie in the range 3100 to 3900.

If we now carry out a trial and obtain, say, 3471 spots, everything is fine: inductive reasoning led to a correct prediction. But suppose that the first trial yields 3958, the second yields 4032, and the third yields 4010. Then something is apparently wrong, since all the results lie outside the "high probability range" 3100 to 3900. There seems to be something in the experimental situation which has a bearing on the outcome but which is not reflected in the given data. Thus, we are led to examine the experimental situation more closely. Perhaps when we do we will discover that the dice are weighted in such a way that only even numbers can come up. The single die probability distribution satisfying Jaynes' principle for the new data is then $P_1(i) = 1/3$ $(i = 2, 4, 6)$, and $P_1(i) = 0$ $(i = 1, 3, 5)$. This distribution leads to a mean value of 4000 for the total number of spots showing on 1000 dice, so that the outcomes 3958, 4032 and 4010 are reasonable on the basis of the new data.

Again, suppose that the first trial yields 1000 spots. This result is so far outside the high probability range that we are immediately led to examine the experimental situation more closely for further relevant data. Perhaps when we do we will discover that the dice are attached to one another in such a way that all must show the same number of spots.

Thus even when highly plausible statistical mechanical predictions turn out to be experimentally wrong, they are still useful because they enable the observer to learn something new. For example, the incorrect statistical mechanical prediction of the black-body radiation curve provided one of the first indications of the inadequacy of classical mechanics.

To summarize: statistical mechanics works because (a) most quantities of

physical interest are highly predictable on the basis of the experimental data, and (b) it is a fact (of science and of everyday life) that highly plausible inductive inferences usually turn out to be true, and furthermore (c) even if a highly plausible prediction turns out to be incorrect, that prediction is still the *best* that could have been made on the basis of the data, and thus an incorrect prediction enables the observer to learn something new about the physical world.

We have not yet discussed the case in which the data does not yield a highly plausible prediction about $g(x)$. Such "indefinite" predictions are common if N is small, and they may arise even if N is large. For example, the phase function $n(R; x)$ (number of particles in the 3-dimensional region R—see Sections 3.4 and 3.5) is indefinite even if N is large, provided only that R is sufficiently small; in this case, statistical uncertainties are experimentally important and are usually called *density fluctuations*.

If the prediction about $g(x)$ is indefinite, then we may proceed in any of several different ways. We may just forget the whole question of predicting $g(x)$. Or we may gather more data, in the hope that a highly plausible prediction will be possible on the basis of the increased data. Or we may make a large number of separate trials, with the same data (i.e. the same initial preparation) on each trial, and check the *average* measured value $\bar{g}$ against $\langle g \rangle_t$, since according to (2.1.47) and (2.1.48) $\bar{g}$ is highly likely to be nearly equal to $\langle g \rangle_t$. Finally, if we cannot gather more data and yet want to make a prediction about $g(x)$ on a *single* trial, we are forced to put our money on the most probable range of values of $g(x)$, realizing that our prediction may be wrong.

3.8 EQUILIBRIUM

A classical mechanical system is in **statistical equilibrium** (or simply **equilibrium**) at time t_1 if $\varrho(x, t_1)$ depends only on the constants of the motion $\phi_1(x), \ldots, \phi_k(x)$:

$$\varrho(x, t_1) = f[\phi_1(x), \ldots, \phi_k(x)]. \tag{3.8.1}$$

By (3.3.16), $f[\phi_1(x), \ldots, \phi_k(x)]$ satisfies Liouville's equation. Hence if the system is in equilibrium at t_1 its probability distribution will remain constant until we either obtain new data or change the Hamiltonian.

Since ϱ is time-independent at equilibrium, all *predictions* about the system must also be time-independent. Note the distinction between mechanical equilibrium (which means that the exact phase-point is time-independent)

and statistical equilibrium. For example, a box of gas may be in statistical equilibrium even though its particles are in rapid motion. Apparently, equilibrium is not so much a property of the system as it is a property of the observer (i.e. of the observer's data).

Returning to an example mentioned in Chapter 1, if a concentration of gaseous U^{235} is diffusing in a box of gaseous U^{238}, it is quite possible for the system to be in equilibrium with respect to an observer who is unable to distinguish the two isotopes but out of equilibrium with respect to an observer who is able to distinguish the two isotopes. It is meaningless to ask whether the system itself is *really* in equilibrium.

According to (3.6.8) and (1), the precise meaning of equilibrium is that the observer's data refers only to constants of the motion.

Most of the equilibrium distributions used in statistical mechanics are dependent only on $H(x)$. The reason is that $H(x)$ is practically the only constant of the motion which is sufficiently simple that it can be measured. Other candidates are the linear and angular momentum, but these are constants of the motion only if $H(x)$ possesses translational and rotational symmetry, which it ordinarily does not.* In fact, as discussed in Section 3.3, a recent result by Sinai[2] indicates that for most systems, there may not even *exist* any constants of the motion other than $H(x)$.

Among all the possible energy-dependent equilibrium distributions, the most popular are the microcanonical distribution $\varrho_{\text{mic}}(x)$ (see (3.6.5)) and the canonical distribution $\varrho_{\text{can}}(x)$ (see (3.6.6)). (Another widely used distribution is the grand canonical distribution, generally regarded as being applicable to systems which can exchange particles with the external world. We will not discuss this distribution.) The microcanonical distribution is generally regarded as being applicable to closed systems (i.e. systems having time-independent Hamiltonians), while ϱ_{can} is regarded as being applicable to systems in weak interaction with a second system, where the interaction is time-dependent but random (i.e. not precisely known).

It is easy to see why ϱ_{mic} is applicable to closed systems. If we can measure only the energy, then our data is of the form $|H(x) - E| < \delta$, where E is a measured number and δ is the experimental error. By (3.6.5), this leads directly to ϱ_{mic}.

The argument leading to ϱ_{can} is not as simple, since systems with time-dependent, random Hamiltonians are not as simple as closed systems with

* For instance, the Hamiltonian for N particles confined to an immovable box does not possess translational or rotational symmetry. The box removes the symmetry.

known Hamiltonians. Let x denote the phase point of the system of interest, let x' denote the phase point of that part of the external world with which the system is in interaction, and assume that the total Hamiltonian of the system-plus-external-world has the form

$$H_T(x, x') = H_0(x) + H_I(x, x') + H_E(x'), \tag{3.8.2}$$

where the interaction term $H_I(x, x')$ is in some sense weak or unimportant compared to $H_0(x)$. The exact Hamiltonian for the system of interest is then

$$H(x, t) = H_0(x) + H_I(x, x'(t)), \tag{3.8.3}$$

where $x'(t)$ is the precise (but unknown!) phase point of the external world. Thus, the system of interest has a time-dependent, random Hamiltonian. Suppose that we measure the "internal Hamiltonian" $H_0(x)$, obtaining the number E_1 (experimental error will be neglected for simplicity). Since the system is not closed we have no reason to believe that further trials will yield the same number E_1, so we proceed to carry out a series of trials, obtaining $E_1, E_2, \ldots, E_n$. If the trials are considered to be statistically independent, and if n is large, then (see (2.1.47) and (2.1.48)) the average value $\bar{E} \equiv \Sigma E_i/n$ is nearly certain to be approximately equal to $\langle H_0(x) \rangle$. It thus seems reasonable to take $\langle H_0(x) \rangle = \bar{E}$ as the given data concerning $H_0(x)$. By (3.6.6), this leads to $\varrho_{\text{can}}(x)$ with $H(x)$ interpreted as the internal Hamiltonian $H_0(x)$. Finally, if the interaction $H_I(x, x')$ is sufficiently weak, then the *first* trial value E_1 should be nearly equal, with high probability, to $\bar{E}$. Thus we can use the data $E_1 = \langle H_0(x) \rangle$, without carrying out a whole series of trials.

The microcanonical distribution contains correlations, regardless of the form of the Hamiltonian. For example, let the data be $|H(x) - E| < \delta$, and let $H(x) = \Sigma \mathbf{p}_i^2/2m + V(\mathbf{q}_1, \ldots, \mathbf{q}_N)$, where V is a non-negative function. The new data $\mathbf{p}_1^2/2m \simeq E$ then implies $\mathbf{p}_i^2/2m \simeq 0$ $(i = 2, \ldots, N)$, so that new information about one particle affects the predictions about all the others.

From (3.6.6) we see that ϱ_{can} is uncorrelated if the particles are non-interacting, and weakly correlated if the particles are weakly interacting.

Often, N is large and the particles are weakly interacting. Then, as will be explained below, ϱ_{mic} and ϱ_{can} are nearly identical so that we may use ϱ_{can} for most purposes even if the system is closed. (It is usually more convenient to work with ϱ_{can} than with ϱ_{mic}, due to the simple analytic properties of the exponential function.) The reason for the similarity of the two distributions is as follows: If the particles are weakly interacting, then ϱ_{can} is weakly correlated. If N is large (one mole, $N \simeq 6 \times 10^{23}$, should be large enough!) then, from Section 3.5, the dispersion in $H(x)$ is small. Thus the Tchebycheff in-

equality (2.1.45) implies that ϱ_{can} is highly concentrated near the energy surface $H(x) = E$, which is just where ϱ_{mic} is concentrated.

We can generalize the above argument: Let the data be $|g_j(x) - G_j| < \delta_j$ $(j = 1, \ldots, k)$, where the $g_j(x)$ are (at least approximately) sum functions of the form (3.4.5), and where the G_j are given numbers and δ_j is the experimental error in measuring $g_j(x)$. By (2.4.7), the appropriate distribution is the "generalized microcanonical distribution" $\varrho_{gmic}(x)$ corresponding to the data. Let $\varrho_{gcan}(x)$ be the "generalized canonical distribution" (2.4.10) corresponding to $\langle g_j(x)\rangle = G_j$ $(j = 1, \ldots, k)$. If N is sufficiently large, the argument given in the preceding paragraph implies that ϱ_{gmic} and ϱ_{gcan} are nearly identical, and thus for most purposes we can use ϱ_{gcan} in place of ϱ_{gmic}. This result explains the great popularity of generalized canonical distributions for many-body systems.

For our particle in a box, the only constant of the motion is the energy $p^2/2m$. Thus all equilibrium distributions are of the form $f(p^2)$. For example, the data $|p| \leqslant b$ leads to the equilibrium distribution

$$\varrho(p, q) = \begin{cases} 1/2bL & (|p| \leqslant b) \\ 0 & (|p| > b). \end{cases}$$

If the data cannot be expressed in terms of p^2, then the distribution will not be an equilibrium distribution. Thus an observer who measures the system with the instruments M_1, M_2, M_3 of Section 3.6 will say that the system is not in equilibrium, whereas another observer who measures the system with only M_3 will say that the system *is* in equilibrium. Once again, we see that equilibrium is not so much a property of the *system* as it is of the *data*.

3.9 THERMAL EQUILIBRIUM AND THE SECOND LAW

For the present, we will restrict the second law and the entropy concept to thermal equilibrium situations; we will give a more general discussion in Chapter 5. In order to discuss the entropy, we must first present a few ideas from thermodynamics.

Consider a many-body system composed of r species of particles, with N_k $(k = 1, 2, \ldots, r)$ indistinguishable particles of the kth species. Let the system be confined to a container of volume V. As discussed in the previous Section, typical equilibrium data for such a system is $\langle H(x)\rangle = E$. We will assume that, in addition to the data $\langle H\rangle = E$, the parameters $V, N_1, N_2, \ldots, N_r$ (required for knowledge of the Hamiltonian function) are known.

The set of particle numbers $N_1, \ldots, N_r$ will be abbreviated $\{N_k\}$. For simplicity, we will assume that V and $\{N_k\}$ are the only parameters which need to be specified in the Hamiltonian. According to (3.6.6), the canonical distribution

$$\varrho_{can}(x) = Z^{-1}(\beta, V, \{N_k\}) \exp[-\beta H(x; V, \{N_k\})], \tag{3.9.1}$$

$$Z(\beta, V, \{N_k\}) = \int \exp[-\beta H(x; V, \{N_k\})]\, dx, \tag{3.9.2}$$

describes the data, where we choose $\beta = \beta(E, V, \{N_k\})$ to satisfy the data $E = \langle H(x; V, \{N_k\})\rangle$, and where we have explicitly included the observable parameters in the notation.

Thermal equilibrium means any situation in which the known data is $\langle H(x)\rangle = E$ plus all parameters needed to specify the Hamiltonian. Thus, the canonical distribution describes thermal equilibrium. **Equilibrium thermodynamics** (or simply **thermodynamics**) means the study of the behavior of the data (or **thermodynamic variables**) $E, V, \{N_k\}$ when a many-body system is taken from one thermal equilibrium state to another.

More precisely, let the system be described at time t_1 by $(E_1, V_1, \{N_{1k}\})$. At some time $\bar{t} > t_1$, let some constraint which had been present at t_1 be changed (for instance, a wall might be moved). We then find that, at time $t_2 \gg \bar{t}$, the only data which remains useful are the new values $(E_2, V_2, \{N_{2k}\})$ of the thermodynamic variables. The system is then said to have relaxed back to thermal equilibrium, and the time for this to occur is called the **relaxation time**. This relaxation is an experimental fact, known as the **zeroth law of thermodynamics**; its explanation in terms of statistical mechanics will be given in Section 5.2. **Thermodynamics** then means the study of the relationship between the final thermal equilibrium state $(E_2, V_2, \{N_{2k}\})$ and the initial state $(E_1, V_1, \{N_{1k}\})$. The idea of thermodynamics is to predict the final state, given the initial state and given the changes in the constraints. Equilibrium thermodynamics makes no attempt to describe what happens *during* the change from the initial to the final state.

We may state the famous **second law of thermodynamics** in the following way: Assume that, due to a change in the constraints, the system evolves from the initial equilibrium state $(E_1, V_1, \{N_{1k}\})$ to the final equilibrium state $(E_2, V_2, \{N_{2k}\})$. Assume, for simplicity, that the change is small. Let δW be the macroscopic work performed on the system during the change:

$$\delta W \equiv \mathbf{F}_{mac} \cdot \delta\mathbf{s}, \tag{3.9.3}$$

where $\mathbf{F}_{mac}$ is the *observable* force acting on the system, and $\delta\mathbf{s}$ is the (small) displacement through which this force is moved in changing the constraints.

Obviously, there may be forces on the system which the observer cannot measure, i.e. which are not included in $\mathbf{F}_{mac}$. Thus, in general $\delta W \neq \delta E$, where $\delta E \equiv E_2 - E_1$. The work

$$\delta Q \equiv \delta E - \delta W^* \tag{3.9.4}$$

done on the system by the non-observable forces is called the **heating** (or "heat added to the system"—an unfortunate terminology). Note that δQ is macroscopically measurable since δE and δW are. The second law then states that there exist thermodynamic variables (i.e. functions of $E, V, \{N_k\}$) $T(E, V, \{N_k\})$ and $S(E, V, \{N_k\})$ such that, to first order in small quantities,

$$\delta S \geqslant \delta Q/T, \tag{3.9.5}$$

with equality if and only if the change in constraints is carried out very slowly. Here, $\delta S \equiv S(E_2, V_2, \{N_{2k}\}) - S(E_1, V_1, \{N_{1k}\})$ is the change in S, and T may mean either $T(E_2, V_2, \{N_{2k}\})$ or $T(E_1, V_1, \{N_{1k}\})$. The functions S and T satisfying the second law are called the **thermodynamic entropy** (or simply **entropy**) and the **temperature**.

The main assertion of the second law is that *there exist* functions $S(E, V, \{N_k\})$ and $T(E, V, \{N_k\})$ having the property (5). Thermodynamics says nothing about the form of these functions for any particular physical system. Given simply that such functions *exist*, practically the entire structure of thermodynamics follows.

It follows from the canonical distribution (1) that functions S and T having the desired property (5) do indeed exist, and in fact are given by

$$S(E, V, \{N_k\}) = -k \int \varrho_{can}(x) \ln [h^{3N} N_1! \cdots N_r! \, \varrho_{can}(x)] \, dx \tag{3.9.6}$$

$$T(E, V, \{N_k\}) = \frac{1}{k\beta(E, V, \{N_k\})}, \tag{3.9.7}$$

where $N = N_1 + \cdots + N_r$ and where k may be any positive constant (k determines only the units of S and T, and is conventionally taken to be Boltzmann's constant). The proof of this statement (i.e. the proof of the second law) will be reserved for Chapter 4, where the quantum statistical version of (5), (6) and (7) will be stated and proved.

It is important to keep in mind that, despite its name, thermodynamics deals only with equilibrium situations and hence is not a dynamical theory

* Equation (4), which says that the total work $\delta Q + \delta W$ equals the increase in energy, is known as the **first law of thermodynamics**. It is a consequence of the principle of conservation of energy.

at all. Thus, statements ascribing a dynamical significance to the second law are incorrect. It is true that the second law implies that, for any adiabatically closed system (i.e. such that $\delta Q = 0$) undergoing the change $(E_1, V_1, \{N_{1k}\})$ to $(E_2, V_2, \{N_{2k}\})$, the final entropy is greater than or equal to the initial entropy:

$$S_2 \geqslant S_1 \quad \text{if} \quad \delta Q = 0. \tag{3.9.8}$$

But this does *not* say that "the entropy always increases with time", or that "the entropy reaches a maximum at equilibrium". In fact, the entropy is defined only at equilibrium and hence cannot "always increase with time;" the entropy has no *meaning* outside of equilibrium, and thus cannot "reach a maximum at equilibrium". The reader is advised to beware of such statements in the literature.

In Chapter 5, we will generalize the entropy concept to non-equilibrium situations. The new "generalized entropy" $S(t)$ *will* then be a dynamical (i.e. time-dependent) quantity. We will find that $S(t)$ *does* agree with the second law of thermodynamics, but that $S(t)$ is *not* necessarily a monotonically increasing function of time. As noted above, this is not a contradiction.

We will now discuss the meaning of the thermodynamic entropy and the significance of the second law.

The entropy depends only on $E, V, \{N_k\}$, and is hence an observable (i.e. a function only of the observable data).

By (6) and (3.6.4), S/k is just the uncertainty associated with ϱ_{can}. But ϱ_{can} itself arises when the data is $(E, V, \{N_k\})$. Thus, aside from the constant k,* *the entropy $S(E, V, \{N_k\})$ is the observer's uncertainty about the phase point x,** when he knows only the macroscopic data $E, V, \{N_k\}$.* Thus, the entropy has a very simple, direct meaning in terms of uncertainty.

The second law says that if a small change of constraints results in heating δQ, then the observer's uncertainty about x increases by more than $\delta Q/T$. In the case of an adiabatic process (i.e. one in which only the observable forces do work, so that $\delta Q = 0$), the uncertainty must increase, or at best remain constant. Thus the second law has a very simple interpretation: *A macroscopic observer, whose data is restricted to the thermodynamic observables, can never gain information about x by manipulating the constraints in an adiabatically closed system; in fact, he will necessarily lose information unless he manipulates the constraints very slowly.*

* Recall from Section 2.3 that the uncertainty is unique only up to a positive multiplicative constant; k merely changes the units of uncertainty.

** More precisely, S is the observer's uncertainty about which phase region of measure $\mathscr{C} = N_1! \cdots N_r! \, h^{3N}$ the phase point lies in; see Section 3.6.

For example, in the free expansion of a gas from volume V_1 to volume V_2, we have $\delta S > 0$. The entropy increase reflects our increased uncertainty about the position $\mathbf{q}_1, \ldots, \mathbf{q}_N$ of the particles: initially each particle was known to be in a box of volume V_1, while finally each particle is known to be in a larger box of volume V_2.

As another example, if we put into thermal contact two boxes of ideal gas, initially at temperatures T_A and T_B, the zeroth law implies that the boxes will attain a final equilibrium state in which each is at the same temperature T. Assuming that the total system is adiabatically enclosed (i.e. $\delta Q = 0$, i.e. $\delta Q_A = -\delta Q_B$, i.e. the only non-observable work is done by one box on the other), the second law implies that $\delta S > 0$. The entropy increase reflects our increased uncertainty about the velocities of the particles. For instance, if we are initially asked to reach into the system and pull out a particle moving faster than the median speed, we have better than a 50% chance of succeeding by reaching into the box having the highest temperature; if we are asked to perform the same task after the new equilibrium state has been attained, we have only a 50% chance of succeeding. Thus we know more about the initial velocities than about the final velocities.

Many physical scientists find it difficult to understand, or interpret, the entropy. It is not difficult to understand an observable like E; since E is just the expected value of the phase function $H(x)$, E has a microscopic interpretation, or mechanical interpretation, as the mechanical energy. Similarly, parameters such as V and $\{N_k\}$ have meaning even at the microscopic level. The entropy, on the other hand, is a purely statistical concept. It has no interpretation as a mechanical phase function, or as the expectation value of a mechanical phase function,* or as a parameter in the Hamiltonian. There is no microscopic, mechanical quantity corresponding to the entropy. Thus one will never understand the entropy if one insists on understanding it in *mechanical* terms. The entropy cannot be understood in mechanical terms because it has no meaning except in connection with a probability distribution; it is meaningless to speak of the entropy of a precise mechanical state x. In order to emphasize the correct statistical meaning of the entropy, it might be helpful to speak of the "entropy of the data" or the "entropy relative to the observer" rather than the entropy of the *system*.

The interpretation of entropy as uncertainty is sometimes objected to on

* It is true that S is, formally, the expectation value of $\phi(x) \equiv -k \ln [h^{3N} N_1! \cdots N_r! \varrho_{can}(x)]$. But $\phi(x)$ is not really a mechanical phase function, since $\varrho_{can}(x(t))$, evaluated at the precise phase point $x(t)$ of the system at time t, does not describe any mechanical property of the system. Instead, the *entire function* $\varrho_{can}(x)$ describes the observer's data.

the grounds that the entropy then becomes a subjective (i.e. relative to the observer) notion, whereas actually the entropy is an observable property of any thermodynamic system. This argument appears to stem from the view (incorrect, in the author's opinion) that every observable quantity must have a purely mechanical interpretation. The entropy is an observable quantity (i.e. a quantity whose value is determined by the *observable* data $E, V, \{N_k\}$) which does *not* have a mechanical interpretation. For any measured values of $E, V, \{N_k\}$, the entropy takes on a perfectly well-defined, measurable value $S(E, V, \{N_k\})$, the same for all observers possessing the same data. Thus the entropy is not really subjective (i.e. relative to the observer), but is instead relative to the observer's *data*. In fact, we might argue that observable quantities such as the entropy are truly objective, whereas such mechanical notions as phase functions and phase points are metaphysical or subjective, since for many-body systems they cannot really be measured.

An interesting point, known as Gibbs' paradox, gives a further illustration of the non-mechanical nature of the entropy. Consider two ideal gases A and B in separate containers. For simplicity, let A and B each have the same temperature T, volume V, and number of particles N. Let the particles of A be distinguishable from the particles of B. It can be shown that, if we allow the two gases to mix by removing the partition between the two containers, an entropy increase

$$S_2 - S_1 = 2kN \ln 2 \quad (A \text{ and } B \text{ distinguishable}) \tag{3.9.9}$$

occurs, where S_1 and S_2 are the total entropies of the equilibrium states before and after removal of the partition. But if the particles of A are indistinguishable from the particles of B, then no observable change occurs when the partition is removed, and so

$$S_2 - S_1 = 0 \quad (A \text{ and } B \text{ indistinguishable}). \tag{3.9.10}$$

Thus an entropy increase $\Delta S = 2kN \ln 2$ (called the entropy of mixing) occurs if and only if A and B are distinguishable.

Equations (9) and (10) are sometimes considered to be paradoxical since, no matter how nearly identical A and B might be, as long as they are the least bit distinguishable, $\Delta S = 2kN \ln 2$; but when they are indistinguishable, ΔS suddenly jumps to zero. If we consider S to be a mechanical quantity, then it does indeed seem unreasonable that ΔS should suddenly jump by the large amount $2kN \ln 2$ when the nature of the system is changed by only a small amount (from *nearly* indistinguishable species to *indistinguishable* species). But if we interpret S as the observer's uncertainty about x, then the

jump seems reasonable since the data undergoes a discontinuity: If A and B are distinguishable then (no matter how nearly identical A and B might be) the initial data is "N particles of type A are in box A, and N particles of type B are in box B"; if A and B are indistinguishable, the initial data is "N particles are in box A and N particles are in box B". In the distinguishable case, we know initially which particles are in which box, while in the indistinguishable case we do not. Thus, in the distinguishable case we lose information when the gases mix, while in the indistinguishable case we do not.

REFERENCES

1. H. Goldstein, *Classical Mechanics* (Addison-Wesley Publishing Company, Inc., Reading, Mass., 1950).
2. Ja. G. Sinai, in *Statistical Mechanics* (Proc. of the IUPAP Meeting, Copenhagen, 1966), edited by T. A. Bak (W. A. Benjamin, Inc., New York, 1967).
3. I. N. Sneddon, *Elements of Partial Differential Equations* (McGraw-Hill Book Company, Inc., New York, 1957).
4. I. Prigogine, *Non-Equilibrium Statistical Mechanics* (John Wiley & Sons, New York, 1962).
5. T. Wu, *Kinetic Equations of Gases and Plasmas* (Addison-Wesley Publishing Company, Reading, Mass., 1966).
6. R. L. Liboff, *Introduction to the Theory of Kinetic Equations* (John Wiley & Sons, New York, 1969).
7. S. A. Rice and P. Gray, *The Statistical Mechanics of Simple Liquids* (John Wiley & Sons, New York, 1965).
8. E. T. Jaynes, IEEE Trans. on Systems Science and Cybernetics *SSC-4*, 227 (1968).
9. E. T. Jaynes, in *Delaware Seminar in the Foundations of Physics*, edited by M. Bunge (Springer-Verlag, Inc., New York, 1967).

CHAPTER 4

Quantum Statistical Mechanics

4.1 THE PROBABILITY ASSIGNMENT

QUANTUM STATISTICAL MECHANICS is the scheme for dealing with quantum mechanical systems (i.e. systems for which quantum effects are important) in terms of information which is incomplete (i.e. less than is permitted by the laws of quantum mechanics). For quantum systems, complete information allows us to determine a quantum state. For a system of spinless particles having f degrees of freedom, a state may be represented by a wave function $\Psi(q_1, \ldots, q_f)$, or in Dirac notation by a vector $|\Psi\rangle$ in the abstract vector space corresponding to the system. Henceforth, we will use the vector space formalism. This formalism is convenient for the same reason that the phase space formalism is convenient in classical statistical mechanics: the formalism encourages us to think of a state as a *single* object (a multi-dimensional vector), so that a probability assignment over a set of states is not difficult to visualize.

A complicating feature of quantum statistics is that even when we have complete information, we still cannot make precise predictions about the outcome of all possible measurements on the system. Even when $|\Psi\rangle$ is known, there remains some uncertainty in the predictions. Now, statistical mechanics deals with incomplete information; for quantum systems, incomplete information means that we cannot even determine $|\Psi\rangle$. Thus, there are *two* sources of uncertainty, and two kinds of probabilistic considerations, in quantum statistics: the statistical mechanical uncertainty concerning which quantum state the system is in, and the quantum mechanical uncertainty which exists even when the state is known.

A complete description of a quantum system is called a **state** or **pure state,** and is represented by a normalized state vector $|\Psi\rangle$. An incomplete description is called a **mixed state**. Note that it is not really the state of the system which is "mixed", since the system is always in some definite quantum state $|\Psi\rangle$; the only thing that is "mixed" is the observer's *data* about the state.*

* It may even be that a precise quantum state $|\Psi\rangle$ is not so much a property of the system itself as it is a property of the observer's *data* about the system. This question belongs to the foundations of quantum theory and will not be discussed here.

As in classical statistics, inductive inference and hence probability theory are called for if the description of the state is incomplete.

It is fairly clear that the appropriate choice of the probability space $\mathscr{S}$ (see Section 2.1) is the set of all vectors $|\Psi\rangle$ representing physically possible states of the system. Thus, $\mathscr{S}$ contains all the normalized vectors $|\Psi\rangle$ in the vector space of the system. (In case the system contains indistinguishable particles, $\mathscr{S}$ contains only the normalized and *properly symmetrized* vectors; see Section 4.4). Hence, $\mathscr{S}$ is a continuum. However, as we will see, quantum statistical probability assignments on $\mathscr{S}$ are generally discrete, so we may take the field of events $\mathbf{F}$ (see Section 2.1) as the collection of all *countable* sets of vectors in $\mathscr{S}$. That is, an event $\mathscr{E}$ means a countable set of vectors of $\mathscr{S}$, and $\mathbf{F}$ consists of all such events $\mathscr{E}$. Finally, a probability assignment P means an assignment of a probability $P(\mathscr{E})$ to every $\mathscr{E} \in \mathbf{F}$.

In science, information is always obtained by making measurements. According to the theory of quantum mechanics, whenever we use an apparatus A to perform a quantum mechanical measurement, the quantum state of the system immediately after the measurement must be one of the eigenvectors $|A_i\rangle$ corresponding to A. That is, letting $\hat{A}$ represent the vector space operator corresponding to A, the state immediately after measurement must be one of the vectors $|A_i\rangle$ satisfying

$$\hat{A}\,|A_i\rangle = A_i\,|A_i\rangle.^* \tag{4.1.1}$$

For most measuring instruments A, the "spectrum" of possible measured values A_i is countable, so that the index i ranges over a discrete set. According to quantum theory, the set $\{|A_i\rangle\}$ of eigenvectors of $\hat{A}$ is complete and orthonormal. That is,

$$\Sigma\,|A_i\rangle\,\langle A_i| = \hat{1} \tag{4.1.2}$$

where $\hat{1}$ is the identity operator, and

$$\langle A_i\,|\,A_j\rangle = \delta_{ij} \tag{4.1.3}$$

where δ_{ij} is the Kronecker delta.

Rather than measuring the system with a single apparatus, we can make a simultaneous measurement with several instruments $A, B, C, \ldots$, provided the corresponding operators $\hat{A}, \hat{B}, \hat{C}, \ldots$ all commute with each other:

$$[\hat{A}, \hat{B}] \equiv \hat{A}\hat{B} - \hat{B}\hat{A} = 0, \quad \text{etc.} \tag{4.1.4}$$

* We will use the notation of (1) throughout this book: the eigenvector of $\hat{A}$ having eigenvalue A_i will be denoted $|A_i\rangle$.

If the set $\hat{A}, \hat{B}, \hat{C}, \ldots$ is a so-called "complete set of commuting operators", and if we know the precise numerical outcome $(A_i, B_j, C_k, \ldots)$ of the measurement, then we know the system to be in the quantum state $|A_i\,B_j\,C_k\,\cdots\rangle$ after the measurement. Thus, we obtain complete quantum mechanical information by precisely measuring a complete set of commuting observables. But if (as is nearly always the case with experiments in the real world) we do *not* measure a complete set of commuting operators and/or we do *not* know the precise numerical outcome of the measurement, then we do not know the quantum state after the measurement. In this case, we must resort to inductive reasoning, i.e. we must describe what we do know about the quantum state by means of a probability distribution over the various states which the system might be in. We call such a probability distribution a **mixture** of states.

Suppose that we carry out a measurement with an apparatus A, but that we do not observe the precise outcome of the experiment, i.e. we do not know which of the possible outcomes A_i actually occurred. We do know, from the general theory of quantum mechanics, that the system is in one of the eigenstates $|A_i\rangle$ after the measurement; our problem is, we don't know precisely which one. Thus, our data may be described by a mixture of the form "the system is in state $|A_i\rangle$ with probability p_i", or more briefly $\{|A_i\rangle, p_i\}$, where the states $|A_i\rangle$ are eigenstates of $\hat{A}$ and hence form a complete and orthonormal set. In Section 4.4, we will present the method of determining the probabilities p_i corresponding to the data.

Note that even though the probability space $\mathscr{S}$ is a continuum, the probability assignment P is *discrete;* i.e. on the basis of the experimental data, the states having non-zero probability form a discrete set. For any $|\Psi\rangle \in \mathscr{S}$, if $|\Psi\rangle$ is *not* one of the eigenstates $|A_i\rangle$ of the instrument used to obtain the data then $P(|\Psi\rangle) = 0$, while if $|\Psi\rangle = |A_i\rangle$ then $P(|\Psi\rangle) = p_i$. Recall that an event $\mathscr{E}$ means a countable set $(|\Psi_1\rangle, |\Psi_2\rangle, \ldots)$ of vectors in $\mathscr{S}$. Thus, for any $\mathscr{E}$, $P(\mathscr{E})$ is the sum of the probabilities of all the eigenvectors $|A_i\rangle \in \mathscr{E}$.

Since the possible states $|A_i\rangle$ are orthonormal, it follows that the mixture $\{|A_i\rangle, p_i\}$ corresponds *uniquely* to the operator

$$\hat{\varrho} \equiv \Sigma\,|A_i\rangle\,p_i\,\langle A_i|; \tag{4.1.5}$$

that is, two mixtures $\{|A_i\rangle, p_i\}$ and $\{|B_j\rangle, p'\}$ are identical if and only if the corresponding operators $\hat{\varrho} = \Sigma\,|A_i\rangle\,p_i\,\langle A_i|$ and $\hat{\varrho}' = \Sigma\,|B_j\rangle\,p'_j\,\langle B_j|$ are the same. Thus, rather than speaking of the mixture $\{|A_i\rangle, p_i\}$, we can speak of the mixture $\hat{\varrho}$. The operator $\hat{\varrho}$ is called the **density operator** corresponding to the data.

The role of $\hat{\varrho}$ in quantum statistics is analogous to the role of $\varrho(x)$ in classical statistics: $\hat{\varrho}$ determines the quantum probability assignment by giving the probability $P(|\Psi\rangle)$ of each state $|\Psi\rangle \in \mathscr{S}$, while $\varrho(x)$ determines the classical probability assignment by giving the probability density at any state $x \in \mathscr{S}$. Such analogies occur often in statistical mechanics, and are helpful in understanding both quantum and classical statistics. The reader is encouraged to find as many analogies as possible between Chapter 3 and the present Chapter.

Quantum mechanics corresponds to data of the form "the state is $|\Psi\rangle$ with probability 1", so that we may think of quantum mechanics as a special case of quantum statistics with density operator $\hat{\varrho} = |\Psi\rangle\langle\Psi|$. Such a pure state density operator is just a projection operator; a more general mixed state (5) is a sum of weighted projection operators, with weighting factors p_i.

Consider the pure quantum state

$$|\Psi\rangle = \Sigma\, a_i\, |A_i\rangle. \tag{4.1.6}$$

The state (6) is sometimes described by the statement that "the system is in the state $|A_i\rangle$ with probability $|a_i|^2$". This description of (6) is misleading, since it leads one to identify the *pure state* (6) with the *mixed state* (5). In fact, if the system is in the state (6) then it is certainly *not* in any one of the states $|A_i\rangle$, since $|A_i\rangle \neq |\Psi\rangle$. The correct way to describe (6) is: "The system is certainly in the state $|\Psi\rangle$, and in this state the probability that a measurement of A will yield the result A_i is $|a_i|^2$". The correct way to describe (5) is: "The system *is* in one of the states $|A_i\rangle$, but we do not know which one, and our knowledge about which state $|A_i\rangle$ the system is in is described by the probabilities p_i". The mixed state (5) contains both statistical mechanical and quantum mechanical uncertainties, while the pure state (6) contains only quantum uncertainties. In Section 4.2 we will give an explicit example illustrating the quantitative and qualitative differences between (5) and (6).

It can be shown that the density operator (5) has the following properties: $\hat{\varrho}$ is Hermitian (since it is a sum of weighted projection operators); $\hat{\varrho}$ has unit trace

$$\mathrm{Tr}\,\hat{\varrho} = 1; \tag{4.1.7}$$

the matrix elements of $\hat{\varrho}$ in the $\hat{A}$-representation (i.e. the representation corresponding to the apparatus A used to obtain the data) are

$$\langle A_i|\,\hat{\varrho}\,|A_j\rangle = p_i\delta_{ij}; \tag{4.1.8}$$

the eigenvectors and eigenvalues of $\hat{\varrho}$ are given by

$$\hat{\varrho}\,|A_i\rangle = p_i\,|A_i\rangle. \tag{4.1.9}$$

Suppose we wish to predict the outcome of a measurement made with the apparatus B, when the data was obtained with A and is described by (5). The precise outcome is not generally predictable, and hence we must base our predictions on the probability distribution $P(B_k)$ of the various possible values B_k of B. According to quantum theory, the B_k are the eigenvalues of the operator $\hat{B}$ corresponding to B, and the probability of observing B_k given that the system is in the state $|A_i\rangle$ is

$$P\,(B_k|\text{ state }|A_i\rangle) = |\langle A_i|\,B_k\rangle|^2. \tag{4.1.10}$$

But, when the data is described by (5), the probability that the system is in the state $|A_i\rangle$ is p_i. Thus (see (2.1.30) and (2.1.9)) the probability distribution for the observable B is

$$\begin{aligned} P\,(B_k) &= \sum_i P\,(B_k|\text{ state }|A_i\rangle)\,P\,(\text{state }|A_i\rangle) \\ &= \sum_i |\langle A_i|\,B_k\rangle|^2\,p_i \\ &= \langle B_k|\sum_i A_i\rangle\,p_i\,\langle A_i|\,B_k\rangle, \\ P\,(B_k) &= \langle B_k|\hat{\varrho}|B_k\rangle. \end{aligned} \tag{4.1.11}$$

Using (11), the expectation value (see (2.1.39)) of B is

$$\begin{aligned} \langle B\rangle &= \sum_k P(B_k)\,B_k = \sum_k \langle B_k|\,\hat{\varrho}\,|B_k\rangle\,B_k \\ &= \sum_k \langle B_k|\,\hat{\varrho}\hat{B}\,|B_k\rangle, \\ \langle B\rangle &= \mathrm{Tr}\,(\hat{\varrho}\hat{B}). \end{aligned} \tag{4.1.12}$$

Thus, probabilities and expectation values have very simple expressions in terms of $\hat{\varrho}$. Equation (12) reduces (as it must, since quantum mechanics is a special case of quantum statistics) to the correct quantum mechanical expression $\langle B\rangle = \langle\Psi|\,\hat{B}\,|\Psi\rangle$ in the pure state case $\hat{\varrho} = |\Psi\rangle\langle\Psi|$.

4.2 ANOTHER SIMPLE EXAMPLE

The simplest quantum system illustrating the notions in this Chapter is a single spin 1/2 particle (such as an electron) in a static external magnetic field. We will assume the particle to be a pure spin system, i.e. we will neglect its motion in space. The beauty of this system is that its abstract vector space has only two dimensions, so that most of the calculations are easy. We will describe the quantum mechanics of the system, and then present a quantum statistical mechanical example.

The normalized vectors $|+\rangle$ and $|-\rangle$ corresponding to "spin up" (spin along the field) and "spin down" (spin opposite to the field) span the vector space of the system. Thus $\mathscr{S}$ consists of all linear combinations

$$|\Psi\rangle = a_1 |+\rangle + a_2 |-\rangle \tag{4.2.1}$$

where a_1 and a_2 may be any complex numbers such that $|a_1|^2 + |a_2|^2 = 1$. Note that $\mathscr{S}$ is a continuum.

Choosing the z axis to be along the field, the Hamiltonian operator is

$$\hat{H} = -\gamma B_0 \hat{s}_z. \tag{4.2.2}$$

Here, γ is the gyromagnetic ratio (ratio of the magnetic moment to the spin angular momentum), B_0 is the magnitude of the external field, and $\hat{s}_z$ is the z component of the spin operator, defined by

$$\hat{s}_z |+\rangle = (1/2)\,\hbar\,|+\rangle, \quad \hat{s}_z |-\rangle = (-1/2)\,\hbar\,|-\rangle, \tag{4.2.3}$$

where $\hbar$ is Planck's constant divided by 2π. It follows from (2) and (3) that $|+\rangle$ and $|-\rangle$ are energy eigenvectors. Schroedinger's equation for the state $|\Psi(t)\rangle$ at time t is

$$\hat{H}\,|\Psi(t)\rangle = i\hbar \frac{d}{dt}\,|\Psi(t)\rangle, \quad i \equiv \sqrt{-1}. \tag{4.2.4}$$

The solution corresponding to the initial state (1) is

$$|\Psi(t)\rangle = a_1 \exp(i\gamma B_0\, t/2)\,|+\rangle + a_2 \exp(-i\gamma B_0\, t/2)\,|-\rangle. \tag{4.2.5}$$

By (3), the matrix corresponding to $\hat{s}_z$ is, in the energy representation,

$$(\hat{s}_z) = \begin{pmatrix} \hbar/2 & 0 \\ 0 & -\hbar/2 \end{pmatrix}. \tag{4.2.6}$$

Another operator which will be needed is the x component of the spin operator, defined by the matrix

$$(\hat{s}_x) = \begin{pmatrix} 0 & \hbar/2 \\ \hbar/2 & 0 \end{pmatrix} \tag{4.2.7}$$

in the energy representation. It follows from (7) that the eigenvectors of $\hat{s}_x$ are

$$\left.\begin{aligned} |+s_x\rangle &= \frac{1}{\sqrt{2}}\,|+\rangle + \frac{1}{\sqrt{2}}\,|-\rangle \\ |-s_x\rangle &= \frac{1}{\sqrt{2}}\,|+\rangle - \frac{1}{\sqrt{2}}\,|-\rangle \end{aligned}\right\} \tag{4.2.8}$$

with eigenvalues $\pm\hbar/2$.

A precise measurement of any spin component s_t determines a precise quantum state, so a single operator $\hat{s}_t$ forms a complete set of commuting observables. If s_t is measured, then the state immediately after measurement is the eigenvector $|s_t\rangle$ corresponding to the measured value.

Suppose that the initial state is given by (1). The probability distributions for s_x and s_z are then

$$\left.\begin{aligned} P(s_x = \hbar/2) &= \tfrac{1}{2}\,|a_1 + a_2|^2 = \tfrac{1}{2} + \mathrm{Re}\,(a_1 a_2^*) \\ P(s_x = -\hbar/2) &= \tfrac{1}{2}\,|a_1 - a_2|^2 = \tfrac{1}{2} - \mathrm{Re}\,(a_1 a_2^*) \\ P(s_z = \hbar/2) &= |a_1|^2 \\ P(s_z = -\hbar/2) &= |a_2|^2. \end{aligned}\right\} \tag{4.2.9}$$

Thus, when the system is in the pure state (1), all predictions about s_x and s_z must be based on the probabilities (9).

All of the foregoing is pure quantum mechanics. Now suppose that a measurement is carried out at $t = 0$ but that, for some reason, the observer doesn't know the precise quantum state of the system immediately after measurement. We then have a quantum statistical mechanical situation.

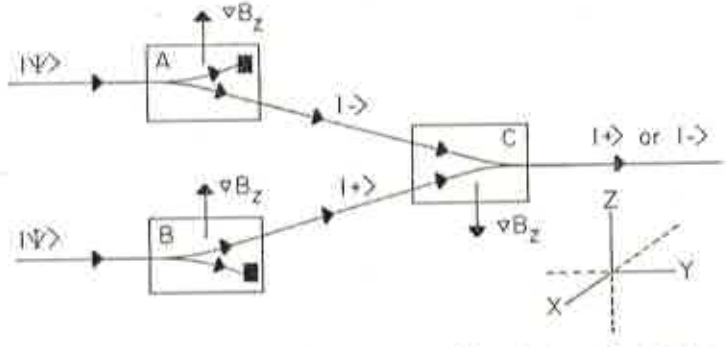

Figure 4.2-1 An experimental arrangement giving rise to a mixed state

For instance, consider the experimental arrangement shown in Figure 1. On the left-hand side are two incoming beams of non-interacting (i.e. widely spaced) spin 1/2 particles moving in the y direction. Each incoming particle is known to be in the same spin state (1). Now let the upper beam be subjected to the Stern-Gerlach apparatus A, and let the lower beam be subjected to the Stern-Gerlach apparatus B.* In both A and B, the field gradient

* See Ref. 1 for an excellent description of idealized Stern-Gerlach experiments.

is in the $+z$ direction; the upper channel (the $s_z = +\hbar/2$ channel) of A is blocked, and the lower channel (the $s_z = -\hbar/2$ channel) of B is blocked. Thus, every particle emerging from A is in the state $|-\rangle$, and every particle emerging from B is in the state $|+\rangle$. Both beams are now sent through a third Stern-Gerlach apparatus C having both channels open and having its field gradient in the $-z$ direction. The sole purpose of C is to straighten out the two beams and bring them together into a single beam.

We are now asked to make predictions about the future behavior of one of the particles emerging from C. *If* we know that the particle came through A, we can assert that the particle is in the state $|-\rangle$; if we know that the particle came through B, we can assert that the particle is in state $|+\rangle$. In principle, we could determine the apparatus (A or B) through which any given particle passed. But suppose that we have in fact *not* determined this. We then know that any single particle emerging from C is either in $|+\rangle$ or $|-\rangle$, but we do not know which. Thus we must base future predictions about the particle on a mixture of the form "$|+\rangle$ with probability p_1, $|-\rangle$ with probability p_2". If the incoming state is given by (1), then it seems reasonable to take $p_1 = |a_1|^2$, $p_2 = |a_2|^2$ (we will justify this choice in Section 4.4). Thus the density operator

$$\hat{\varrho} = |+\rangle\,|a_1|^2\,\langle +| + |-\rangle\,|a_2|^2\,\langle -| \tag{4.2.10}$$

describes a single outgoing particle.

The two incoming "beams" could consist of just a *single particle*, without affecting the discussion in any way. That is, if a single particle in the known initial state $|\Psi\rangle$ is sent through A or B and then emerges from C, and if the observer does not know which apparatus (A or B) the particle went through, then the observer's state of knowledge about the particle is described by the mixed state (10).

Since $|a_1|^2 + |a_2|^2 = 1$, the density operator (10) has unit trace, in agreement with (4.1.7). The density operator (10) is diagonalized in the s_z representation, in agreement with (4.1.8). Using (4.1.11), we find that predictions about s_x and s_z in the mixed state (10) must be based on the probabilities

$$\begin{aligned} P(s_x = \hbar/2) &= 1/2 \\ P(s_x = -\hbar/2) &= 1/2 \\ P(s_z = \hbar/2) &= |a_1|^2 \\ P(s_z = -\hbar/2) &= |a_2|^2. \end{aligned} \tag{4.2.11}$$

Comparing (9) and (11), we see that there are similarities between the pure state (1) and mixed state (10) (predictions about s_z are the same), but there are also differences (predictions about s_x are different). There is greater uncertainty in the mixed state than in the pure state, since s_x is entirely unpredictable in the mixed state but is more-or-less predictable (depending on the values of a_1 and a_2) in the pure state. This is a reflection of the fact that (10) contains not only the quantum uncertainty contained in (1), but also the statistical mechanical uncertainty arising from lack of knowledge of the quantum state.

As an example of (4.1.12), the density operator (10) leads to

$$\langle s_x \rangle = \mathrm{Tr}\,(\hat{\varrho}\hat{s}_x) = 0.$$

Compare this with the result

$$\langle s_x \rangle = \langle \Psi |\, \hat{s}_x\, | \Psi \rangle = \mathrm{Re}\,(a_1 a_2^*)$$

for the pure state (1).

4.3 DYNAMICS OF THE PROBABILITY ASSIGNMENT

As in classical statistics, quantum statistical data gathered at one time should be useful for making predictions about other times. The information is propagated by the dynamics of the system.

Corresponding to the state $|\Psi_0\rangle$ at t_0, denote the state at time t by $\hat{T}(t, t_0)\,|\Psi_0\rangle$; thus, $|\Psi_0\rangle$ evolves into $|\Psi\rangle = \hat{T}(t, t_0)\,|\Psi_0\rangle$ during t_0 to t. The basic dynamical condition on the system is Schroedinger's equation:

$$\hat{H}\,|\Psi\rangle = i\hbar \frac{d}{dt}\,|\Psi\rangle. \tag{4.3.1}$$

The quantum mechanical motion defines a one-to-one transformation of the vector space of the system.

We can translate the derivation of (3.3.3) into quantum language, to obtain the basic condition which the dynamics imposes on the probability assignment:

> If $|\Psi_0\rangle$ evolves into $|\Psi_1\rangle$ during t_0 to t_1, then the probability of $|\Psi_1\rangle$ at t_1 equals the probability of $|\Psi_0\rangle$ at t_0. (4.3.2)

Thus, if the data is described at $t = 0$ by the mixture $\{|\Psi_j\rangle, p_j\}$, then the appropriate mixture at t is $\{\hat{T}(t, 0)\,|\Psi_j\rangle, p_j\}$, i.e. we assign the *same* prob-

abilities p_j to the states $\hat{T}(t,0)|\Psi_j\rangle$ as were assigned to the corresponding states $|\Psi_j\rangle$ at $t=0$. In terms of the density operator, the initial mixed state

$$\hat{\varrho}(0) = \Sigma |\Psi_j\rangle\, p_j\, \langle\Psi_j| \tag{4.3.3}$$

evolves into

$$\hat{\varrho}(t) = \Sigma |\Psi_j(t)\rangle\, p_j\, \langle\Psi_j(t)| \tag{4.3.4}$$

where $|\Psi_j(t)\rangle \equiv \hat{T}(t,0)|\Psi_j\rangle$.

Using (1) and (4), we can derive the following evolution equation for $\hat{\varrho}$:

$$\frac{d}{dt}\hat{\varrho}(t) - \frac{i}{\hbar}[\hat{\varrho}(t), \hat{H}] = 0. \tag{4.3.5}$$

Equation (5) is the analogue of the Liouville equation (3.3.13); we will call it the **quantum Liouville equation**.

The matrix elements for $\hat{\varrho}(t)$ in energy representation have a simple time dependence:

$$\langle H_j|\hat{\varrho}(t)|H_k\rangle = \langle H_j|\hat{\varrho}(0)|H_k\rangle \exp\left[\frac{i}{\hbar}(H_k - H_j)t\right]. \tag{4.3.6}$$

Proof: By (4) and (3),

$$\begin{aligned}\hat{\varrho}(t) &= \sum_j |\Psi_j(t)\rangle\, p_j\, \langle\Psi_j(t)| \\ &= \sum_j \hat{T}(t,0)|\Psi_j\rangle\, p_j\, \langle\Psi_j|\hat{T}^\dagger(t,0) \\ &= \hat{T}(t,0)\,\hat{\varrho}(0)\,\hat{T}^\dagger(t,0),\end{aligned}$$

where $\hat{T}^\dagger$ means the adjoint of $\hat{T}$. Thus

$$\langle H_j|\hat{\varrho}(t)|H_k\rangle = \langle H_j|\hat{T}(t,0)\,\hat{\varrho}(0)\,\hat{T}^\dagger(t,0)|H_k\rangle.$$

But Schroedinger's equation (1) implies

$$\hat{T}(t,0)|H_k\rangle = \exp\left(\frac{-i}{\hbar}H_k t\right)|H_k\rangle.$$

Thus

$$\langle H_j|\hat{\varrho}(t)|H_k\rangle = \langle H_j|\exp\left(\frac{-i}{\hbar}H_j t\right)\hat{\varrho}(0)\exp\left(\frac{i}{\hbar}H_k t\right)|H_k\rangle,$$

which gives (6).

As in the classical case (see Section 3.3), the dynamics described by (2) are appropriate only for translating data at t_0 into predictions at t_1. If, on the other hand, we obtain new data at some time $\bar{t}$ between t_0 and t_1, then we must revise the probability assignment at $\bar{t}$ to reflect the new data. Thus $\hat{\varrho}(t)$

obeys the quantum Liouville equation at all times except times $\bar{t}$ when new data is obtained. At such times $\bar{t}$, $\hat{\varrho}(t)$ undergoes a sudden jump or "collapse". This behavior appears to be simply a generalization of the so-called "collapse of the wave function" which occurs in quantum mechanics: when a system in the quantum state $|\Psi(t)\rangle$ is subjected to a measurement of some observable A at time $\bar{t}$, the state "collapses" at $\bar{t}$ from $|\Psi(t)\rangle$ to the eigenstate $|A_i\rangle$ corresponding to the measured value A_i of A. We have seen that the quantum statistical explanation for the collapse of a *mixed* state $\hat{\varrho}$ is simply that new data is obtained upon measurement, and the probability assignment $\hat{\varrho}$ must be changed to reflect the new data. A basic question in quantum mechanics, known as the "measurement problem", is whether the same explanation is appropriate for the collapse of a *pure* quantum state, or whether on the other hand the collapse of a pure state can be explained mechanically in terms of the interaction between the system and the measuring apparatus during the measurement. We will not consider this question in this book.

In the case of our spin 1/2 particle in a static magnetic field (see Section 4.2), $\hat{\varrho}(t)$ has the four matrix elements $\langle\pm|\hat{\varrho}(t)|\pm\rangle$ in the energy representation. By (4), the matrix elements representing $\hat{\varrho}(t)$ when $\hat{\varrho}(0)$ is given by (3) are

$$\langle\pm|\hat{\varrho}(t)|\pm\rangle = \sum_j \langle\pm|\Psi_j(t)\rangle\, p_j\, \langle\Psi_j(t)|\pm\rangle. \tag{4.3.7}$$

For any given $\hat{\varrho}(0)$, we can find these elements explicitly by using (4.2.5). The matrix elements may also be found directly from (6).

For example, if $\hat{\varrho}(0)$ is given by (4.2.10) then

$$\hat{\varrho}(t) = \hat{\varrho}(0),$$

i.e. $\hat{\varrho}(t)$ doesn't change with time. Thus all predictions are constant in time.

As a more interesting example, suppose that the initial data is obtained by measuring s_x, but that the precise outcome ($s_x = \hbar/2$ or $s_x = -\hbar/2$) isn't known. Then (see Section 4.1) $\hat{\varrho}(0)$ has the form

$$\hat{\varrho}(0) = |+s_x > p < +s_x| + |-s_x > q < -s_x|, \tag{4.3.8}$$

where $p + q = 1$ and where $|\pm s_x\rangle$ are the eigenvectors of $\hat{s}_x$, given by (4.2.8). By (4.2.5), the corresponding states at time t are

$$|\pm s_x, t\rangle = \frac{1}{\sqrt{2}}\exp\left(\frac{i\gamma B_0 t}{2}\right)|+\rangle \pm \frac{1}{\sqrt{2}}\exp\left(\frac{-i\gamma B_0 t}{2}\right)|-\rangle. \tag{4.3.9}$$

By (4),

$$\hat{\varrho}(t) = |+s_x, t\rangle\, p\, \langle+s_x, t| + |-s_x, t\rangle\, q\, \langle-s_x, t|. \tag{4.3.10}$$

Equations (7), (9) and (10) imply that the density matrix in energy representation is

$$(\hat{\varrho}(t)) = 1/2 \begin{pmatrix} 1 & (p-q)\exp(i\gamma B_0 t) \\ (p-q)\exp(-i\gamma B_0 t) & 1 \end{pmatrix} \tag{4.3.11}$$

(We could also obtain this from (6)). From (4.1.12) along with (4.2.2), (4.2.6) and (4.2.7), it follows that

$$\begin{aligned} \langle s_x \rangle_t &= \tfrac{1}{2}\hbar\,(p-q)\cos(\gamma B_0 t) \\ \langle s_z \rangle_t &= 0 \\ \langle H \rangle_t &= 0. \end{aligned} \tag{4.3.12}$$

The expectation values of s_z and H are constant, as they should be since s_z and H are constants of the motion (i.e. their operators commute with $\hat{H}$). The expectation value of s_x is oscillatory. Equations (4.2.6) and (4.2.7) imply that $\hat{s}_x^2 = \hat{s}_z^2 = (\hbar^2/4)\,\hat{1}$, so that $\langle s_x^2 \rangle = \langle s_z^2 \rangle = \hbar^2/4$. Using (12), it follows that the variances in s_x and s_z are

$$\begin{aligned} \sigma^2(s_x) &= (\hbar^2/4)\,[1-(p-q)^2\cos^2(\gamma B_0 t)], \\ \sigma^2(s_z) &= \hbar^2/4. \end{aligned} \tag{4.3.13}$$

The component s_z is always highly unpredictable, since its variance is always $\hbar^2/4$; the variance of s_x oscillates between $(\hbar^2/4)\,[1-(p-q)^2]$ and $\hbar^2/4$, so that the predictability of s_x is oscillatory. In the pure state case $p = 1$ or $p = 0$, in which the outcome of the initial s_z measurement is precisely known, the variance $\sigma^2(s_x)$ oscillates between 0 and $\hbar^2/4$. At the other extreme, if $p = q = 1/2$ then $\sigma^2(s_x) = \hbar^2/4$ for all t.

Contrast the oscillatory behavior in (12) and (13) with the relaxation to equilibrium found in our simple classical statistical example (see for instance Figure (3.4-5)). This contrast between quantum statistics and classical statistics is not simply a peculiarity of the two examples chosen, but is characteristic of the two theories, and appears to be one of the fundamental differences between quantum and classical statistics. This difference may be traced back to the continuous nature of the possible outcomes in a classical measurement versus the discrete nature of the possible outcomes in a quantum measurement. We will discuss this question further in Chapter 5.

4.4 THE INITIAL DISTRIBUTION

Suppose that we obtain data D as the result of a measurement of an observable A (or several commuting observables). If D determines a specific quantum state, then there is no need for quantum statistics. But if D does not determine a specific state then, according to Section 4.1, the initial information is described by a statement of the form "the system is in state $|A_i\rangle$ with probability p_i" where the $|A_i\rangle$ are the eigenvectors of $\hat{A}$ and where the p_i have yet to be determined. As in classical statistics, we determine the p_i through Jaynes' principle.

For quantum systems, two situations are common:

Case 1. Nothing is known about the state $|\Psi\rangle$ just prior to the measurement of A, and the measurement yields data D.

The opposite of this case is:

Case 2. The precise state $|\Psi\rangle$ is known just prior to the measurement of A, and there is no data about the outcome of the measurement.

Case 1 is the situation studied in classical statistics, in which we have no data prior to the initial measurement, and the measurement yields partial information. In Case 2, we have complete data prior to the measurement, but no information about the outcome of the measurement. Case 2 is not interesting in classical statistics, because if the classical state just prior to the measurement is known to be the phase point x_0, and if (as may be assumed in classical mechanics) the measurement does not disturb the state of the system, then the precise state just after the measurement is known to be x_0 even though the outcome of the measurement is not observed.

We will now discuss Case 1, and will return to Case 2 shortly. Recall (see Section 4.1) that an *outcome* means a state $|\Psi\rangle \in \mathscr{S}$, where $\mathscr{S}$ is the set of all vectors representing physically possible states of the system. In Case 1, our knowledge about the outcome is described by the probability assignment $\{|A_i\rangle, p_i\}$, where the p_i have yet to be determined. Thus the uncertainty about the outcome $|\Psi\rangle$ is the uncertainty in the probability distribution $(p_1, p_2, \ldots)$. According to (2.3.28), the expression for this uncertainty depends on the distribution P^m and P^0 representing the maximum information and the prior information (prior to the measurement). We will now determine the distributions P^m and P^0.

There is no prior information in Case 1, so P^0 must represent complete ignorance. It is fairly clear that "complete ignorance" is symmetric with

respect to the possible outcomes $\{|A_i\rangle\}$, so that for an n-dimensional vector space the prior probabilities are $p_i^0 = 1/n$. A more formal argument is: Let "complete ignorance" be represented by the density operator $\hat{\varrho}_0$. As in Chapter 3 (see (3.6.2)), $\hat{\varrho}_0$ must commute with every possible Hamiltonian on the vector space of the system. It follows that $\hat{\varrho}_0 = n^{-1}\hat{1}$. Expanding in terms of the A_i,

$$\hat{\varrho}_0 = \Sigma\, |A_i\rangle\, n^{-1}\, \langle A_i|,$$

from which $p_i^0 = 1/n$.

The distribution P^m should correspond to knowledge of the precise quantum state. However, if the system consists of N identical particles, then (as discussed for classical systems in Section 3.6) a question arises as to whether *maximum* information can determine a precise vector $|\Psi\rangle$, since the vector space contains $N!$ different vectors differing only by particle exchange whereas even maximum information cannot distinguish between different particles (and hence cannot distinguish between these $N!$ vectors). Recalling that the probability space $\mathscr{S}$ contains only those vectors representing *physically possible* states, this question is taken care of by the quantum mechanical exchange-symmetry rule. This rule (which is presumed to be familiar to the reader) says that for a system of identical particles, the physically possible states are symmetric (for Bose particles) or anti-symmetric (for Fermi particles) under particle exchange. It follows that, if $|\Psi\rangle$ and $|\Psi'\rangle$ are two physically possible vectors (i.e. two vectors contained in $\mathscr{S}$) differing only by particle permutation, then $|\Psi\rangle$ and $|\Psi'\rangle$ are in fact the *same* vector (except possibly for a difference in sign): $|\Psi\rangle = \pm\,|\Psi'\rangle$. Thus, maximum information does indeed determine a single quantum state $|\Psi\rangle \in \mathscr{S}$.

Thus, in Case 1, P^0 and P^m are precisely as specified in (2.3.30), and Jaynes' principle takes the following form:

> **Jaynes' Principle** for quantum statistical mechanics. If nothing is known about the state $|\Psi\rangle$ prior to the initial measurement, and if data D is obtained by measuring some observable (or set of commuting observables) A, then the initial probability assignment is $\{|A_i\rangle, p_i\}$, where the $|A_i\rangle$ are the eigenvectors of $\hat{A}$ and where the p_i maximize the uncertainty
>
> $$U\,(t=0) = -\Sigma\, p_i \ln p_i$$
>
> subject to D. (4.4.1)

The expression in (1) for the uncertainty may be written in terms of $\hat{\varrho}$ as

$$U\,[\hat{\varrho}] = -\mathrm{Tr}\,\hat{\varrho}\ln\hat{\varrho}.^{*} \tag{4.4.2}$$

Proof: Evaluating the trace in the $\hat{A}$ representation,

$$-\mathrm{Tr}\,\hat{\varrho}\ln\hat{\varrho} = -\sum_{ij}\langle A_i|\,\hat{\varrho}\,|A_j\rangle\,\langle A_j|\ln\hat{\varrho}\,|A_i\rangle.$$

But $\langle A_i|\,\hat{\varrho}\,|A_j\rangle = \delta_{ij}p_i$, and (see footnote below) $\langle A_j|\ln\hat{\varrho}\,|A_i\rangle = \delta_{ij}\ln p_i$, so

$$-\mathrm{Tr}\,\hat{\varrho}\ln\hat{\varrho} = -\Sigma\, p_i \ln p_i,$$

which is precisely the expression given in (1) for the uncertainty.

Two important consequences of Jaynes' principle are:

> Let the energy H be measured, and let the data be $E_1 \leqslant H \leqslant E_2$ where E_1 and E_2 are given. Then
>
> $$\hat{\varrho}\,(t=0) = \Sigma\, |H_i\rangle\, n^{-1}\, \langle H_i|,$$
>
> where the sum is over all $|H_i\rangle$ such that $E_1 \leqslant H_i \leqslant E_2$, and where n is the number of such eigenvectors $|H_i\rangle$. (4.4.3)

> Let the energy H be measured, and let the data be $\langle H\rangle = E$ where E is given. Then
>
> $$\hat{\varrho}\,(t=0) = Z^{-1}\exp(-\beta\hat{H})$$
>
> where $Z = \mathrm{Tr}\exp(-\beta\hat{H}) \equiv Z(\beta)$ and β is chosen to satisfy the data. (4.4.4)

These results follow by maximizing $-\Sigma p_i \ln p_i$ subject to the data, where $p_i = P(|H_i\rangle)$. In (3) the maximizing probabilities are $p_i = n^{-1}$ $(E_1 \leqslant H_i \leqslant E_2)$. In (4) $p_i = \exp(-\beta H_i)$. The density operators (3) and (4) are called the **microcanonical density operator** and the **canonical density operator**. They are the quantum analogues of (3.6.5) and (3.6.6).

According to Section 4.1, if data is obtained by simultaneous measurement of $(A, B, \ldots)$ then the corresponding density operator is diagonal in any

---

* A word about functions of operators (for instance, $\ln\hat{\varrho}$): An operator $\hat{A}$ is "functionally dependent" on $\hat{B}$ if $\hat{A}$ is diagonalized in every representation diagonalizing $\hat{B}$. In this case, the statement $\hat{A} = f(\hat{B})$, where $f(x)$ is some ordinary (i.e. numerical) function, means that the eigenvalues of $\hat{A}$ and $\hat{B}$ are related by $A_i = f(B_i)$. In case $f(x)$ is expandable in a power series $\Sigma\, a_n (x - x_0)^n$, this definition implies that $f(\hat{B}) = \Sigma\, a_n (\hat{B} - x_0)^n$.

representation which simultaneously diagonalizes $(\hat{A}, \hat{B}, \ldots)$. According to the footnote on page 107, this implies that $\hat{\varrho}$ is functionally dependent on $\hat{A}, \hat{B}, \ldots$:

> If the data is obtained by simultaneous measurement of $(A, B, \ldots)$, then $\hat{\varrho}$ has the form $\hat{\varrho} = f(\hat{A}, \hat{B}, \ldots)$. (4.4.5)

This result is the quantum analogue of (3.6.8). Note that $(\hat{A}, \hat{B}, \ldots)$ commute, since $(A, B, \ldots)$ are simultaneously measurable.

We will now discuss Case 2, in which we know $|\Psi\rangle$ just prior to the measurement of some observable A but do not know the outcome of the measurement. Expand $|\Psi\rangle$ in terms of the $|A_i\rangle$:

$$|\Psi\rangle = \Sigma\, a_i\, |A_i\rangle .$$

Quantum theory says that, *before* the measurement, our knowledge about what state the system will be in *after* the measurement is described by the statement "$|A_i\rangle$ has probability $p_i^0 = |a_i|^2$". The measurement itself yields no new data, since we do not know the outcome of the measurement. Thus the probabilities p_i describing our knowledge immediately after the measurement must maximize the uncertainty $U(P; P^m; P^0)$ (see (2.3.28)) where P^0 is specified by $p_i^0 = |a_i|^2$ and where the only condition on P is $\Sigma p_i = 1$. By (2.3.28), maximization of U is equivalent to minimization of

$$I(P; P^0) = \Sigma\, p_i \ln (p_i/p_i^0).$$

By (2.3.15), $I(P; P^0)$ is minimized (for fixed P^0) when $p_i = p_i^0$. Thus, in Case 2 the density operator is

$$\hat{\varrho}\,(t = 0) = \Sigma\, |A_i\rangle\, |a_i|^2\, \langle A_i| \tag{4.4.6}$$

where $a_i \equiv \langle \Psi | A_i \rangle$ and $|\Psi\rangle$ means the state just before measurement. The density operator (6) is appropriate only if we do not know the outcome of the measurement of A. If, on the other hand, we know that the outcome is A_i, then $\hat{\varrho}\,(t = 0)$ is the pure state $|A_i\rangle\, \langle A_i|$.

For example, let our simple spin 1/2 system be in the state (4.2.1) just prior to a measurement of s_z, and assume that the result of the s_z measurement is not known. (In Section 2 we described a fairly realistic situation in which the outcome might be unknown.) The system is then described by the density operator (4.2.10) just after the measurement.

In Case 2, information is actually lost when the measurement is made. This is rather unusual, since one ordinarily hopes to *gain* information by measurement. The trouble is that the measurement disturbs the system, but

since (in Case 2) we do not observe the outcome, we have no information about the effect of this disturbance on the system. Quantitatively, the uncertainty before the measurement is

$$-\mathrm{Tr}\, \hat{\varrho} \ln \hat{\varrho} = -\mathrm{Tr}\, (|\Psi\rangle\, \langle \Psi| \ln |\Psi\rangle\, \langle \Psi|) = 0.$$

Using (6), the uncertainty immediately after measurement is

$$-\mathrm{Tr}\, \hat{\varrho} \ln \hat{\varrho} = -\Sigma\, |a_i|^2 \ln |a_i|^2 \geqslant 0,$$

with equality only if $|a_i|^2 = 1$ for some a_i, i.e. only if the prior state $|\Psi\rangle$ is one of the $|A_i\rangle$.

It may be worth noting that the information loss which occurs in Case 2 does not have anything to do with the second law of thermodynamics, since equilibrium thermodynamics does not apply to Case 2.

4.5 EQUILIBRIUM

Practically everything about equilibrium in Sections 3.8 and 3.9 applies with only minor modifications to quantum statistics, so our discussion of equilibrium quantum statistics will be brief.

A quantum mechanical **constant of the motion** is an observable whose operator $\hat{A}$ commutes with the Hamiltonian:

$$[\hat{H}, \hat{A}] = 0. \tag{4.5.1}$$

It follows that $\langle A \rangle$ is constant, and the eigenvectors $|A_i\rangle$ are stationary states (i.e. $\exp(i\omega t)\, |A_i\rangle$ satisfies Schroedinger's equation provided ω is chosen appropriately).

A quantum system is in **statistical equilibrium** if $\hat{\varrho}$ depends only on the constants of the motion:

$$\hat{\varrho} = f\,[\hat{A}, \hat{B}, \ldots] \tag{4.5.2}$$

where $\hat{A}, \hat{B}, \ldots$ are constants of the motion. By (4.4.5), the system is in equilibrium whenever the data refers only to constants of the motion. The operator (2) commutes with $\hat{H}$, so (4.3.5) implies $d\hat{\varrho}/dt = 0$. Thus, the system remains in equilibrium until either new data is obtained or $\hat{H}$ is changed. Predictions are time-independent at equilibrium.

The microcanonical and canonical density operators $\hat{\varrho}_{\mathrm{mic}}$ and $\hat{\varrho}_{\mathrm{can}}$ (see (4.4.3) and (4.4.4)) are functions of $\hat{H}$ and hence are equilibrium distributions; $\hat{\varrho}_{\mathrm{mic}}$ is applicable to closed systems, while $\hat{\varrho}_{\mathrm{can}}$ is applicable to systems which exchange energy with the environment. If the number of particles N is large,

then for most purposes $\hat{\varrho}_{\text{mic}}$ and $\hat{\varrho}_{\text{can}}$ yield macroscopically indistinguishable predictions, and hence we may use $\hat{\varrho}_{\text{can}}$ whether the system is open or closed.

If we obtain data by measurement of observables $G^{(1)}, G^{(2)}, \ldots$ whose operators are (at least approximately) a sum of single-particle operators, and if N is large, then the data is described reasonably well by the generalized canonical operator

$$\hat{\varrho}\,(t=0) = Z^{-1} \exp\left[-\Sigma\, \alpha_i\, \hat{G}^{(i)}\right]. \tag{4.5.3}$$

See Section 3.8 for the proof of the corresponding statement in classical statistics.

For our simple spin 1/2 particle in an external magnetic field, $\hat{H}$ is the only functionally independent constant of the motion, i.e. any other constant of the motion is a function of $\hat{H}$. Thus $\hat{\varrho}$ represents equilibrium if and only if $\hat{\varrho} = f(\hat{H})$, i.e. if and only if $\hat{\varrho}$ has the form

$$\hat{\varrho} = |+\rangle\, p\, \langle +| + |-\rangle\, q\, \langle -| \tag{4.5.4}$$

(see the footnote on p. 107 regarding functions of operators).

Equilibrium thermodynamics is based on the canonical density operator

$$\hat{\varrho}_{\text{can}} = Z^{-1}(\beta, V, \{N_k\}) \exp\left[-\beta \hat{H}\,(V, \{N_k\})\right], \tag{4.5.5}$$

$$Z(\beta, V, \{N_k\}) = \operatorname{Tr} \exp\left[-\beta \hat{H}\,(V, \{N_k\})\right]; \tag{4.5.6}$$

the symbols $\beta, V, \{N_k\}$ have been defined on p. 86. A system described by $\hat{\varrho}_{\text{can}}$ is said to be in **thermal equilibrium**.

We will now state and prove the second law of thermodynamics for thermal equilibrium states: Consider a small (but otherwise arbitrary) change in the constraints on a system, resulting in an evolution from the initial equilibrium state $(E, V, \{N_k\})$ to the final equilibrium state $(E + \delta E, V + \delta V, \{N_k + \delta N_k\})$ The second law asserts that there exist thermodynamic variables (i.e. functions of E, V, and $\{N_k\}$) $T(E, V, \{N_k\})$ and $S(E, V, \{N_k\})$ such that, to first order in small quantities, the change in S due to the change in state satisfies

$$\delta S \geqslant \delta Q/T. \tag{4.5.7}$$

Here,

$$\delta Q \equiv \delta E - \delta W \tag{4.5.8}$$

is called the **heating** done on the system during the process, and δW is the macroscopically observable work done on the system during the change in constraints.

Proof of the second law: We will show that the thermodynamic variables defined by

$$S(E, V, \{N_k\}) = -k \operatorname{Tr}(\hat{\varrho}_{\text{can}} \ln \hat{\varrho}_{\text{can}}), \tag{4.5.9}$$

$$T(E, V, \{N_k\}) = \frac{1}{k\beta\,(E, V, \{N_k\})}, \tag{4.5.10}$$

satisfy (7). For simplicity, we will assume that only one particle species is present, so that the observables are (E, V, N); the proof may be easily extended to r species.

By (5) and (9),

$$\begin{aligned} S(E, V, N) &= -k \operatorname{Tr}\left\{\hat{\varrho}_{\text{can}} \ln\left[\frac{\exp(-\beta\hat{H})}{Z}\right]\right\} \\ &= -k \operatorname{Tr}\left[\hat{\varrho}_{\text{can}}(-\beta\hat{H} - \ln Z)\right] \\ &= k\beta E + k \ln Z(\beta, V, N). \end{aligned} \tag{4.5.11}$$

Taking the differential of (11), the change in S due to the change in state is, to first order in small quantities,

$$\begin{aligned} \delta S = {} & k\beta\, \delta E + kE\delta\beta + k \frac{\partial \ln Z(\beta, V, N)}{\partial \beta} \delta\beta \\ & + k \frac{\partial \ln Z(\beta, V, N)}{\partial V} \delta V + k \frac{\partial \ln Z(\beta, V, N)}{\partial N} \delta N. \end{aligned} \tag{4.5.12}$$

From (6),

$$\frac{\partial \ln Z(\beta, V, N)}{\partial \beta} = \frac{1}{Z} \frac{\partial}{\partial \beta} \operatorname{Tr}[\exp(-\beta\hat{H})] = -\frac{1}{Z} \operatorname{Tr}[\hat{H} \exp(-\beta\hat{H})] = -E. \tag{4.5.13}$$

Thus the second and third terms on the right-hand side of (12) cancel, and (12) becomes

$$\delta S = k\beta\, \delta E + k \frac{\partial \ln Z(\beta, V, N)}{\partial V} \delta V + k \frac{\partial \ln Z(\beta, V, N)}{\partial N} \delta N. \tag{4.5.14}$$

Equation (14) gives the change in S for any small change $(\delta E, \delta V, \delta N)$.

Let the change in constraints be initiated at time t_1 and terminated at t_2 (note that the time t_1 to t_2 may be long, even though the change in constraints is assumed to be small). The macroscopic work done on the system during the change in constraints may be written,

$$\delta W = \int_{t_1}^{t_2} \dot{W}(t)\, dt,$$

where $\dot{W}(t)$ is the rate of doing macroscopic work at time t (note that $\dot{W}(t)$ is a small quantity). Since, for many-body systems, macroscopic quantities are equal (with high probability) to the expectation values of corresponding microscopic quantities,

$$\delta W = \int_{t_1}^{t_2} \langle \hat{\dot{W}}_{\text{mic}} \rangle_t \, dt, \tag{4.5.15}$$

where

$$\langle \hat{\dot{W}}_{\text{mic}} \rangle_t \equiv \text{Tr}\,[\hat{\varrho}(t)\, \hat{\dot{W}}_{\text{mic}}] \tag{4.5.16}$$

is the expected rate of doing work at time t.

Now let us assume, for a moment, that the process is **quasi-static,** i.e. that the change in constraints is carried out infinitely slowly. The zeroth law of thermodynamics (i.e. the fact that systems relax to thermal equilibrium) then implies that the system remains in equilibrium throughout the change. Thus, at any time t during the process, the system is represented by the canonical density operator (5), where β, V, and N depend on t. But β, V, and N change by only a small amount during the entire process. Hence, to lowest order in small quantities, $\langle \hat{\dot{W}}_{\text{mic}} \rangle_t$ in (15) may be replaced by $\langle \hat{\dot{W}}_{\text{mic}} \rangle_{t_1}$, where $\langle\ \rangle_{t_1}$ represents an expectation value taken over the *initial* canonical density operator. Equation (15) may then be written

$$\delta W = \int_{t_1}^{t_2} \langle \hat{\dot{W}}_{\text{mic}} \rangle_{t_1} \, dt = \left\langle \int_{t_1}^{t_2} \hat{\dot{W}}_{\text{mic}} \, dt \right\rangle_{t_1} = \langle \widehat{\delta W}_{\text{mic}} \rangle_{t_1} \tag{4.5.17}$$

where $\widehat{\delta W}_{\text{mic}}$ is the operator representing the total microscopic work done on the system during the change in constraints. But

$$\widehat{\delta W}_{\text{mic}} = \hat{H}(t_2) - \hat{H}(t_1) = \frac{\partial \hat{H}}{\partial V}\, \delta V + \frac{\partial \hat{H}}{\partial N}\, \delta N, \tag{4.5.18}$$

where the second step follows from the fact that $\hat{H}$ depends parametrically only on V and N. Equations (17) and (18) imply that the macroscopic work done in a quasi-static process is

$$\delta W = \text{Tr}\left[\hat{\varrho}_{\text{can}}(t_1)\left(\frac{\partial \hat{H}}{\partial V}\, \delta V + \frac{\partial \hat{H}}{\partial N}\, \delta N\right)\right] = \left\langle \frac{\partial \hat{H}}{\partial V} \right\rangle_{t_1} \delta V + \left\langle \frac{\partial \hat{H}}{\partial N} \right\rangle_{t_1} \delta N. \tag{4.5.19}$$

A derivation analogous to (13) shows that

$$\frac{\partial \ln Z(\beta, V, N)}{\partial V} = -\beta \left\langle \frac{\partial \hat{H}}{\partial V} \right\rangle_{t_1}$$
$$\frac{\partial \ln Z(\beta, V, N)}{\partial N} = -\beta \left\langle \frac{\partial \hat{H}}{\partial N} \right\rangle_{t_1}. \tag{4.5.20}$$

Combining (14), (19) and (20), the change in S due to a quasi-static change in constraints is

$$\delta S = k\beta\, \delta E - k\beta\, \delta W = \delta Q/T, \tag{4.5.21}$$

where the second step follows from (8) and (10). Thus, the equality in (7) holds for quasi-static processes.

Now consider a non-quasi-static process connecting the states (E, V, N) and $(E + \delta E, V + \delta V, N + \delta N)$. In this case, the change in S is still given by (14), but the work done is no longer given by (19). The work done on the system in a non-quasi-static process connecting two given states must be greater than the work done in a quasi-static process, since in the non-quasi-static case the applied macroscopic forces must not only alter the constraints, but must also accelerate the system. It follows that the heating $\delta Q = \delta E - \delta W$ is *less* in the non-quasi-static case than in the quasi-static case. Thus, the inequality in (7) holds for any non-quasi-static process. This completes the statistical mechanical proof of the second law.

Thus we have proven the second law by actually exhibiting thermodynamic variables (9) and (10) having the desired property (7). The interpretation of the second law and of the entropy have been given in Section 3.9.

Gibbs' paradox (i.e. the jump from (3.9.9) to (3.9.10) in the entropy of mixing) is sometimes considered to be less paradoxical in quantum statistics than in classical statistics. The classical explanation (see Section 3.9) for the apparent paradox is that our *information* changes radically when we alter the system from nearly indistinguishable species to completely indistinguishable species, and hence we expect the entropy of mixing to change radically. In quantum statistics, the jump in the entropy of mixing still occurs, but now the explanation seems different: the jump occurs because the set of physically possible states (satisfying the exchange-symmetry rule) is altered drastically when we alter the system from nearly indistinguishable species to completely indistinguishable species. Thus, the classical explanation appears to be non-mechanical, whereas the quantum explanation appears to be given in terms of a purely mechanical principle (the exchange-symmetry rule).

In the author's opinion, the above-described distinction between the classical and quantum explanations of Gibbs' paradox is merely verbal. In fact, entirely within classical mechanics, we could invent an "exchange-symmetry rule" according to which the only physically admissible classical states, for a system of N indistinguishable particles, are symmetric under particle exchange. A single classical state would then no longer be represented by a single point x_0 in phase space, but would instead be represented by $N!$ points differing only by particle permutations. A single point (or outcome) in the probability space $\mathscr{S}_\mu$ of physically possible states would then correspond to $N!$ points in phase space, and all integrals would be taken over $\mathscr{S}_\mu$. Indeed, this approach is sometimes used in classical statistics (although not ordinarily in classical *mechanics*); $\mathscr{S}_\mu$ is called the "generic phase space" and a point in $\mathscr{S}_\mu$ is called the "generic phase". If this approach is used, the classical expression for the entropy will no longer contain the factor $N!$ and one can argue that Gibbs' paradox has a purely mechanical explanation in terms of the new classical exchange-symmetry rule. The only reason that such a symmetry rule is not used in classical mechanics is that it is not convenient to think of a classical state as a symmetrized combination of points in phase space.

Thus, the classical and quantum explanations of Gibbs' paradox are fundamentally the same, and the classical explanation is fully as natural as the quantum explanation.

Although Gibbs' paradox is fully explainable in terms of classical statistical mechanics, there are important aspects of the quantum exchange-symmetry rule which have no classical analogue. From symmetry considerations, it follows (in both the classical and quantum cases) that all *predictions* about systems of indistinguishable particles must be symmetric under particle exchange. In classical mechanics, it follows that the state itself must be symmetric. But in quantum mechanics, predictions (e.g. $\langle A\rangle = \langle\Psi|\hat{A}|\Psi\rangle$) are unchanged when $|\Psi\rangle$ is multiplied by $\exp(i\phi)$, so it follows only that $|\Psi\rangle$ must be symmetric *to within a phase factor*. More precisely, if $|\Psi\rangle$ and $|\Psi'\rangle$ differ only by particle exchange, then it follows by symmetry that $|\Psi'\rangle = \exp(i\phi)|\Psi\rangle$, where ϕ is a real angle. The basic physical assertion of the quantum exchange-symmetry rule is then that $\phi = 0$ for systems of Bose particles (with spin 0, 1, 2, ...) and $\phi = \pi$ for systems of Fermi particles (with spin (1/2, 3/2, ...). It appears that we cannot deduce this assertion from symmetry arguments of the kind used to explain Gibbs' paradox. The assertion that $\phi = 0$ or π affects quantum statistical predictions in a very basic way, since it determines which states are counted as physically possible and hence

determines the structure of the probability space $\mathscr{S}$. The resulting statistical predictions are called "Bose statistics" for Bose systems and "Fermi statistics" for Fermi systems. No analogous pair of statistics exists in classical statistical mechanics, and thus the exchange-symmetry rule is responsible for one of the basic differences between classical and quantum statistics.

The **third law of thermodynamics** states that the equilibrium thermodynamic entropy approaches zero as $T \to 0$:

$$\lim_{T\to 0} S(T) = 0. \tag{4.5.22}$$

We can derive the third law from quantum statistics: By (5), (9) and (10),

$$S(T) = -k \Sigma P(|H_i\rangle) \ln P(|H_i\rangle), \tag{4.5.23}$$

where

$$P(|H_i\rangle) = Z^{-1}(T) \exp(-H_i/kT). \tag{4.5.24}$$

In the limit $T \to 0$, (24) becomes

$$\lim_{T\to 0} P(|H_i\rangle) = \begin{cases} n^{-1} & (H_i = H_{\min}) \\ 0 & (H_i > H_{\min}) \end{cases} \tag{4.5.25}$$

where $H_{\min}$ is the smallest energy eigenvalue, and where n is the degeneracy of $H_{\min}$.* Assuming that the sum (23) is uniformly convergent on some interval containing $T = 0$ (so that the limit may be interchanged with the sum), (25) implies

$$\lim_{T\to 0} S(T) = -k \Sigma \lim_{T\to 0} [P(|H_i\rangle) \ln P(|H_i\rangle)] = k \ln n. \tag{4.5.26}$$

If $H_{\min}$ is non-degenerate, then $n = 1$ and (26) is the third law. If n is not too large, then (26) implies the "approximate" form of the third law: $S \to S_0 \simeq 0$ as $T \to 0$.

The third law has a natural interpretation in terms of our interpretation (see Section 3.9) of the entropy. At $T = 0$, the system is in the state $|H_{\min}\rangle$ with probability 1, so (assuming that $H_{\min}$ is non-degenerate) the observer knows the precise quantum state of the system and the uncertainty S is zero. That is, the third law says that if a thermodynamic observer (who can measure only E, V, and $\{N_k\}$) finds that $T(E, V, \{N_k\}) = 0$, then he knows the precise quantum state of the system.

It can be shown that the classical expression for the entropy goes to $-\infty$ as $T \to 0$. Thus we can derive the third law from quantum statistics but not

* We must assume that $H_{\min}$ exists, i.e. that the energy spectrum has a lower bound.

from classical statistics. This is an example of the general rule that quantum effects are important in thermal equilibrium at low temperatures. The reason is that as T becomes small, statistical mechanical uncertainties (i.e. the observer's uncertainty about which mechanical state the system is in) become small; in fact, this is precisely what is asserted by the third law! Thus quantum uncertainties (i.e. the uncertainties which remain even when the precise quantum state $|\Psi\rangle$ is known) begin to predominate over statistical mechanical uncertainties as $T \to 0$. Since classical statistics describes only the statistical mechanical uncertainties and neglects the quantum uncertainties, we expect that classical statistics will be wrong at low temperatures.

REFERENCES

1. R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman Lectures on Physics*, Vol. III (Addison-Wesley Publishing Company, Reading, Mass., 1965).

CHAPTER 5

Irreversibility

5.1 THE REVERSIBILITY OF MECHANICS

One of the most obvious features of the physical world is the irreversibility of most processes: waterfalls always flow downhill; two thermally connected systems evolve toward a common temperature; boxes slide down inclined planes while heating the plane and the box; a cue ball hits a triangular arrangement of 15 pool balls and scatters them. In each case, the time-reversed process is mechanically possible but is not observed in nature: the water molecules could initially be moving upward with sufficient velocity to carry them from the bottom to the top of the waterfall; the molecules of the cold system could spontaneously lose energy to the molecules of the hot system; the box could slide up the inclined plane while cooling the box and the plane; the pool balls could find themselves moving in such a way that 15 of them come together in a triangular arrangement with the total energy being transferred to the cue ball. One of the tasks of statistical mechanics is to explain why these reversed processes, which are mechanically possible, are not in fact observed in the real world. This is called the **irreversibility problem.**

Since the reversed processes are mechanically *possible*, the most we can hope to do (short of rejecting mechanics!) is to prove that they are somehow *improbable*. It follows that probabilistic considerations are an essential part of any explanation of irreversibility, and that any attempt at a purely mechanical (i.e. non-probabilistic) explanation is bound to fail. That is, we can explain irreversibility only in terms of *statistical* mechanics.

The irreversibility concept has given rise to lots of controversies during the past century. Many of these controversies boil down to simple disagreements over the meanings of words. In order to avoid this situation, let us give a few formal definitions. We will say that a theory (such as mechanics, or statistical mechanics) is **reversible** if it is both time-symmetric and non-relaxing; we will say that a theory is **irreversible** if it is both non-time-symmetric and relaxing. We must now define time-symmetry and relaxation. A theory is **time-symmetric** if, for any permitted time-sequence of states $\ldots, A, B, C, \ldots$ the time-reversed sequence $\ldots, C, B, A, \ldots$ is also permitted.

A theory exhibits **relaxation** if any permitted time-sequence of states possesses a limit as $t \to \infty$, i.e. if states "settle down" as $t \to \infty$. Note that non-time-symmetry and relaxation are independent concepts, and that we will call a theory "irreversible" only if it is *both* non-time-symmetric and relaxing. Another useful concept is **almost-periodicity**: A theory is **almost periodic** if the time-dependent state of the system keeps returning arbitrarily near to the initial state (we will sharpen this concept below). Note that almost-periodicity implies non-relaxation.

We will find, in this Section, that classical and quantum mechanics are time-symmetric and almost periodic, and hence reversible. In Section 5.2, we will find that classical statistical mechanics is non-time-symmetric and relaxing, and hence irreversible. In Section 5.3, we find the rather surprising result that quantum statistical mechanics is non-time-symmetric but almost periodic. Hence quantum statistics is neither entirely irreversible nor entirely reversible; it is a little bit of each.

The following theorems show the time-symmetry of mechanics:

Time-symmetry of Classical Mechanics. Let the phase path corresponding to the initial condition (q_0, p_0) be $(q(t), p(t))$. The phase path corresponding to the initial condition $(q_0, -p_0)$ is then $(q(-t), -p(-t))$. (5.1.1)

Time-symmetry of Quantum Mechanics. Let the solution of Schroedinger's equation corresponding to the initial condition $\Psi_0(q)$ be $\Psi(q, t)$. Then the solution corresponding to the initial condition $\Psi_0^*(q)$ is $\Psi^*(q, -t)$. (5.1.2)

Here, $q = (\mathbf{q}_1, \ldots, \mathbf{q}_N)$, $p = (\mathbf{p}_1, \ldots, \mathbf{p}_N)$ and (for simplicity) we assume the quantum system to consist of N spinless particles and we use the wave function formulation of quantum mechanics. The classical theorem follows from the fact that if $q(t)$ and $p(t)$ satisfy Hamilton's equations, then so do $q(-t)$ and $-p(-t)$. The quantum version follows from the fact that if $\Psi(q, t)$ satisfies Schroedinger's equation, then so does $\Psi^*(q, -t)$. Either version says that if a certain sequence of events is possible, then the time-reversed sequence is also possible (where the momenta are of course reversed in the time-reversed sequence). Thus mechanics is time-symmetric: it is mechanically possible for waterfalls to flow uphill, etc. Note that time-symmetry holds even in the limit $N \to \infty$.

The following recurrence theorems show that mechanics is non-relaxing.

Almost-Periodicity* of Classical Mechanics. Let the initial phase-point x_0 be such that the phase path $X(t)$ is confined to a finite region of phase space (i.e. confined within a sphere of finite radius in the $6N$-dimensional phase space). Then, for any $\varepsilon > 0$, there exist infinitely many times τ (spread over the whole range $-\infty$ to ∞) such that $|X(\tau) - x_0| < \varepsilon$ (i.e. $X(\tau)$ is inside a $6N$-dimensional sphere of radius ε with center at x_0). (5.1.3)

Almost-Periodicity of Quantum Mechanics. Let the initial state $|\Psi(0)\rangle$ be expressible as a linear superposition of a discrete set of energy eigenstates. Then, for any $\varepsilon > 0$, there exist infinitely many times τ (spread over the whole range $-\infty$ to ∞) such that $||\Psi(\tau)\rangle - |\Psi(0)\rangle| < \varepsilon$. Here, the norm $||\Psi\rangle|$ is defined by $||\Psi\rangle|^2 = \langle\Psi|\Psi\rangle$. (5.1.4)

We will not prove these theorems here. Poincaré first proved the classical theorem; Ref. 2 gives a good discussion. The quantum theorem was first proved by Ono[3] and independently by Bocchieri and Loinger.[4] These theorems imply that mechanical observables (classical phase functions or quantum expectation values) keep returning (to any desired degree of approximation) to their original values, and hence that relaxation to a final, asymptotic state cannot occur.

Almost-periodicity breaks down in the limit $N \to \infty$, since the recurrence time τ may become infinite in this limit. (In terms of the classical recurrence theorem, $X(t)$ is no longer confined to a finite region of phase space as $N \to \infty$; in terms of the quantum recurrence theorem, the set of energy eigenstates becomes a continuum as $N \to \infty$). It may then seem that irreversibility can be explained within a purely mechanical framework, simply by allowing N to become large. There are at least three things wrong with this opinion: (1) Irreversibility is actually observed even when N is finite (for example, $N = 16$ in the foregoing pool ball example). (2) Letting $N \to \infty$ doesn't change the fact that mechanical systems are time-symmetric. (3) Letting $N \to \infty$ may remove almost-periodicity, but it does not lead to relaxation;

---

* H. Bohr[1] introduced the term *almost periodic* in connection with nearly recurrent functions $f(t)$. The use of this term in (5.1.3) and (5.1.4) is essentially the same as Bohr's usage.

neither finite nor infinite mechanical systems settle down to static mechanical states as $t \to \infty$. Thus, the mechanical evolution of both finite and infinite systems is time-symmetric and non-relaxing, and hence completely reversible.

5.2 IRREVERSIBILITY AND CLASSICAL STATISTICAL MECHANICS

There are two important features of statistical mechanics which are not found in mechanics: statistical mechanics deals with a phase space probability distribution while mechanics deals with a single phase point; the statistical initial condition must obey Jaynes' principle, while there are no general restrictions on the mechanical initial condition. It will be shown in this Section that these two non-mechanical ideas, together with the fact that experimental data never refers to the future, are sufficient to explain irreversibility in terms of classical statistics.

First, consider the implications of the fact that statistical mechanics deals with a probability distribution satisfying Liouville's equation. Defining $q = (\mathbf{q}_1, \ldots, \mathbf{q}_N)$, $p = (\mathbf{p}_1, \ldots, \mathbf{p}_N)$, if $\varrho(q, p, t)$ satisfies Liouville's equation then so does $\varrho(q, -p, -t)$. Thus we have:

> **Time-symmetry of Liouville's Equation.** Let the initial distribution $\varrho(q, p, 0)$ evolve, during 0 to t, into $\varrho(q, p, t)$. Then the initial distribution $\varrho(q, -p, 0)$ evolves, during 0 to t, into $\varrho(q, -p, -t)$. (5.2.1)

Thus Liouville's equation is time-symmetric: if a certain sequence of predictions (e.g. a certain sequence of expectation values) is consistent with Liouville's equation, then so is the time-reversed sequence (where the momenta are reversed in the time-reversed sequence). For instance, in the case of two systems in thermal contact, for every solution predicting that (with high probability) the hot system loses energy to the cold system, there is another solution predicting that (with high probability) the cold system loses energy to the hot system. However, this does not necessarily mean that *statistical mechanics* is time-symmetric, since statistical mechanics is based not only on Liouville's equation but also on Jaynes' principle. In fact, it will turn out that those initial conditions which are consistent with Jaynes' principle lead to the usual irreversible predictions, while the time-reversed predictions arise from initial conditions which are not consistent with Jaynes' principle for any experimentally obtainable data.

We now ask: does Liouville's equation imply almost-periodic predictions? It is easy to see that the answer is "no". For instance, in the particle in a box

example we found in Section 3.3 that $P(\{q < L/2\}; t) \to 0.5$ and $P(\{p > 0\}; t) \to 0.5$ as $t \to \infty$; in Section 3.4 we found that $\langle q \rangle_t \to L/2$ and $\langle p \rangle_t \to 0$ as $t \to \infty$. Even though the mechanical motion of this one-particle system is exactly periodic, probabilities and expectation values settle down to constants as $t \to \infty$. Thus, there do exist finite systems for which the statistical mechanical predictions are not almost-periodic, and hence the almost-periodicity of classical mechanics does not carry over into classical statistical mechanics.

Let us be more precise. We will say that a function $f(x, t)$ **converges strongly** to $f_0(x)$ if, for almost all x (i.e. all x except possibly for a set of Lebesgue measure* zero),

$$\lim_{t \to \infty} f(x, t) = f_0(x). \tag{5.2.2}$$

We will say that $f(x, t)$ **converges weakly** to $f_0(x)$ if, for every Lebesgue integrable* $g(x)$,

$$\lim_{t \to \infty} \int g(x) f(x, t)\, dx = \int g(x) f_0(x)\, dx, \tag{5.2.3}$$

where the integral is over the entire range of x. It may be shown[5] that strong convergence implies weak convergence; the converse is not true. It may be shown[6] that condition (3) for weak convergence is equivalent to the condition that, for every Lebesgue measurable region R of positive measure,

$$\lim_{t \to \infty} \int_R f(x, t)\, dx = \int_R f_0(x)\, dx. \tag{5.2.4}$$

We will say that a differential equation for functions $f(x, t)$ exhibits **relaxation** if every Lebesgue integrable initial condition $f(x, 0)$ leads to a solution $f(x, t)$ which is weakly convergent.

For the particle in a box, solutions of Liouville's equation are not strongly convergent. In fact, with the initial condition (3.2.1), $\lim_{t \to \infty} \varrho(q, p, t)$ does not exist; at the fixed point (q, p), ϱ just continues to alternate between $2/bL$ and 0 with period $2Lm/p$. However, solutions are weakly convergent: the "stretching" or "phase-mixing" of $\varrho(x, t)$ shown in Figure 3.3-2 clearly implies that, for any region R of positive measure (i.e. having positive area in Figure 3.3-2),

$$\lim_{t \to \infty} \int_R \varrho(x, t)\, dx = \int_R \varrho_{eq}(x)\, dx \tag{5.2.5}$$

---

* Regarding *Lebesgue measure* and *Lebesgue integrable*, see the footnotes on pages 18 and 54.

where (assuming the initial condition (3.2.1)

$$\varrho_{eq}(x) = \begin{cases} 1/2bL & (|p| \leqslant b) \\ 0 & (|p| > b). \end{cases} \tag{5.2.6}$$

Thus ϱ converges weakly to ϱ_{eq}. It then follows from (3) that expectation values of all Lebesgue integrable* phase functions $g(x)$ converge:

$$\lim_{t\to\infty} \langle g(x)\rangle_t = \int g(x)\, \varrho_{eq}(x)\, dx. \tag{5.2.7}$$

Note that (6) is an equilibrium distribution, since it depends only on the constant of the motion $|p|$.

It seems clear that, for the particle in a box, the phase-mixing phenomenon (see Figure 3.3-2) will occur whenever $\varrho\,(x, 0)$ is not concentrated on a set of isolated points or curves. [This qualification is important. For example, discrete initial distributions

$$\varrho\,(x, 0) = \Sigma\, p_k\, \delta\,(x - x_k) \tag{5.2.8}$$

will *not* lead to phase-mixing (see Appendix C). Neither will initial distributions

$$\varrho\,(q, p, 0) = f(q)\, \delta\,(p - p_0)$$

concentrated on the line $p = p_0$. On the other hand, initial distributions which are spread over a continuum of p values should phase-mix, since points having different momenta move with different periods and hence the distribution must "stretch".] Thus, for the particle in a box, *Liouville's equation exhibits relaxation,* i.e. all Lebesgue integrable solutions converge weakly to equilibrium distributions, and hence all statistical predictions relax to equilibrium.

For the particle in a box, relaxation results from the fact that the period of mechanical motion is a continuous and non-constant function of the energy of the initial phase point. Since initial points having different energies move with different periods, the distribution stretches. The same behavior occurs for any bound** one-dimensional system whose period* is energy-dependent.

---

* It is worth noting that the Dirac delta function is *not* a Lebesgue integrable function, and hence $\langle \delta\,(q - q_0)\, \delta\,(p - p_0)\rangle_t$, $\langle \delta\,(q - q_0)\rangle_t$, etc. don't necessarily converge.

** A bound system means a system for which each phase path can be entirely enclosed within a finite sphere in the $2f$-dimensional phase space. The radius of the sphere may be different for different paths, i.e. we do not require bound systems to have finite phase spaces.

* Any bound one-dimensional conservative system is necessarily periodic.

Now, it appears that the only bound one-dimensional system whose period is *not* energy dependent is the perfect harmonic oscillator.* Thus, Liouville's equation exhibits relaxation for all bound one-dimensional systems (i.e. systems having one degree of freedom) except for the perfect harmonic oscillator.

Continuing to generalize, Liouville's equation exhibits relaxation for any system of non-interacting particles, provided only that the particles are not harmonic oscillators, since each particle undergoes phase-mixing in its own one-particle phase space. Finally, any interacting many-body system which is equivalent, under an appropriate canonical transformation, to a non-interacting system (i.e. whose transformed Hamiltonian is "separable") exhibits relaxation, provided only that the equivalent system is not a set of harmonic oscillators.

Now, physically realistic many-body systems don't ordinarily have separable Hamiltonians, nor can they be canonically transformed to a system having a separable Hamiltonian. Thus, we must ask whether Liouville's equation exhibits relaxation for these more general systems. The answer appears to be "yes", although a rigorous, general proof has never been given. The argument for the relaxation of Liouville's equation runs as follows: Relaxation is due to the phase-mixing of the distribution function. Phase-mixing occurs whenever points initially close together in phase space do not remain close together as the points move along their respective phase paths. In mathematical terms: Liouville's equation exhibits relaxation if and only if the solutions to Hamilton's equations (more precisely, all solutions except possibly a set of Lebesgue-measure zero) are **non-stable**.** Unfortunately, not much is known about the stability of mechanical systems. We have seen above that the only one-dimensional system whose solutions are stable is the perfect harmonic oscillator. The only non-trivial (i.e. non-separable) mechanical system about which much is known is the restricted three-body problem (three gravitationally interacting particles, where the mass of the third particle is so small that its motion does not affect the other two particles). Even in this problem,

---

* The author is not aware of any rigorous proof of this statement. However, all non-harmonic potentials (such as the slightly anharmonic oscillator, and the square well) used in physically realistic problems seem to have energy dependent periods.

** A solution $X\,(t|x_0)$ is **stable** if, for any $\varepsilon > 0$, there exists $\delta > 0$ such that every initial point x_1 satisfying $|x_1 - x_0| < \delta$ leads to a solution $X\,(t|x_1)$ satisfying $|X\,(t|x_1) - X(t|x_0)| < \varepsilon$ for all $t\,(-\infty < t < \infty)$. In other words, $X\,(t|x_0)$ is a stable solution if $X\,(t|x_1)$ may be kept arbitrarily close to $X\,(t|x_0)$ for all time, simply by choosing x_1 sufficiently close to x_0.

which is probably the simplest of all non-trivial many-body problems, not a single solution is known to be stable[7]. It is strongly suspected that non-separable many-body systems are non-stable except in very exceptional cases[7].

Thus we have (subject to the qualifications of the preceding paragraph):

> **Relaxation of Classical Statistical Mechanics.** If the initial probability distribution is Lebesgue integrable, then probabilities and expectation values settle down to equilibrium as $t \to \infty$:
>
> $$P(x \in R; t) \to \int_R \varrho_{eq}(x)\, dx,$$
>
> $$\langle g(x) \rangle_t \to \int g(x)\, \varrho_{eq}(x)\, dx.$$
>
> Here, R is any phase region of positive Lebesgue measure, $g(x)$ is any Lebesgue integrable function, and $\varrho_{eq}(x)$ is an equilibrium distribution. (5.2.9)

Relaxation of Liouville's equation means that, despite the almost-periodicity of the mechanical motion, statistical mechanical *predictions* settle down to equilibrium. For example, suppose that a gas containing N particles is initially confined by a partition to the left-hand half of a box, and that at $t = 0$ the partition is removed. If we are asked to predict the number of particles n in the left-hand side after, say, ten minutes, based only on the information that $n = N$ at $t = 0$, our prediction is $\langle n \rangle = N/2$. Furthermore, if N is large (larger than, say, 100) then this prediction is highly likely to be nearly correct (see Section 3.5). That is, it is highly likely that $n \simeq N/2$ at $t = 10$ minutes, or at any other time which is long compared to the relaxation time (the relaxation time is the time for $\langle n \rangle$ to "relax" to approximately $N/2$). All of this is true despite the fact that, according to the almost-periodicity of mechanics, there is a *possibility* that at any particular time t, all the particles have spontaneously returned to the left-hand half of the box.

Thus the fact that statistical mechanics deals with a probability distribution rather than with a single phase point is sufficient to introduce the approach to equilibrium into the theory, although it is not sufficient to introduce non-time-symmetry.

Before discussing how non-time-symmetry enters into statistical mechanics, let us pause to consider the character of the distribution $\varrho_{eq}(x)$ to which the

exact distribution relaxes (in the sense of weak convergence) for long times. We can determine the asymptotic distribution directly from $\varrho(x, 0)$ and the constants of the motion $\phi_1(x), \ldots, \phi_m(x)$, without solving Liouville's equation. This is due to the facts that $\varrho_{eq}(x)$ and $\varrho(x, 0)$ must yield the same reduced distributions for the $\phi_i(x)$, and $\varrho_{eq}(x)$ must depend *only* on the $\phi_i(x)$ (see (3.8.1)). These two conditions completely determine $\varrho_{eq}(x)$ in terms of $\varrho(x, 0)$ and the $\phi_i(x)$.

Note that $\varrho(x, t)$ does not necessarily relax to *thermal* equilibrium, since $\varrho_{eq}(x)$ is not necessarily the canonical distribution (3.9.1). According to the preceding paragraph, ϱ relaxes to thermal equilibrium if the following conditions are satisfied:

(a) The energy is the only constant of the motion;
(b) At $t = 0$, the reduced distribution for energy is the canonical energy distribution, i.e. for any E and for small ΔE,

$$\int_{E < H(x) < E + \Delta E} \varrho(x, 0)\, dx = \frac{\exp(-\beta E)}{Z} \int_{E < H(x) < E + \Delta E} dx. \quad (5.2.10)$$

We will discuss these two conditions.

The question of the number and character of the constants of the motion is a deep and difficult problem of classical mechanics. Since $\varrho_{eq}(x)$ must be Lebesgue integrable, we are concerned here only with the number of *Lebesgue integrable* constants of the motion. The question is: is the energy the only such function? For any system whose Hamiltonian is the sum of two or more parts, each part depending on a different set of particles, the answer is "no", since each partial Hamiltonian is conserved. Thus a system of non-interacting particles will not in general relax to thermal equilibrium. But it seems reasonable (or at least conceivable) that interactions (no matter how small) between the particles could "destroy" all the one-particle constants of the motion so that the only ones remaining are the seven standard constants of the motion: energy, the three components of the total momentum, and the three components of the total angular momentum. But if the system is confined by external forces, the momentum integral is destroyed; and if the confining forces are not all directed toward a single force center, then the angular momentum integral is destroyed.

Thus, it appears plausible that, for most interacting many-body systems, the energy is the only Lebesgue integrable constant of the motion, so that condition (a) holds. In fact, Sinai[8] has verified this conjecture for the special

case of N hard spheres confined to a box, where N may be any integer greater than or equal to 2.

There seems to be no particular reason to suppose that $\varrho(x, 0)$ will satisfy condition (b). If (a) is satisfied but (b) is not, then $\varrho(x, t)$ will relax (in the sense of weak convergence) to an equilibrium distribution which is dependent only on the energy but which is non-canonical. Now, recall that throughout this book, all systems have been assumed to be *closed*, i.e. to have non-random, time-independent Hamiltonians. It is possible that the introduction of a random interaction or of a time-dependent interaction between the system and its environment could re-arrange the energy distribution, and drive the system to *thermal* equilibrium. We conclude, finally, that the relaxation of $\varrho(x, t)$ to an energy-dependent equilibrium distribution is a general feature of statistical mechanics for closed, finite systems, but that interactions with the environment are probably responsible for finally driving $\varrho(x, t)$ to the *canonical* distribution.

Note that, even in cases where $\varrho(x, t)$ does not relax weakly to $\varrho_{\text{can}}(x)$, it is still possible for statistical predictions about a limited number of observables to approach the canonical predictions as $t \to \infty$. In a specific example to be studied in Section 5.6, it will be found that the expected total momentum and expected center of mass converge to the canonical predictions, even though $\varrho(x, t)$ does not relax weakly to $\varrho_{\text{can}}(x)$.

We will now examine the question of the non-time-symmetry of statistical mechanics. We have seen that Liouville's equation, considered without reference to other principles of statistical mechanics, is not sufficient to introduce non-time-symmetry. We will find that Jaynes' principle, plus the fact that experimental data may refer to the present or the past but not to the future, implies that statistical mechanical predictions are non-time-symmetric. In this Section, we will show this only by looking at a few simple examples. We will give the general mathematical formulation of this result in Section 5.4.

Let the system be an N-body gas in a box, and suppose that the measuring instrument measures the precise number of particles n in the left-hand half of the box. At $t = t_0$, let the following data be given:

(i) At the present instant t_0, all the particles are in the left-hand half. That is, $n(t_0) = N$.

If we apply Jaynes' principle to this data, we will get the usual irreversible predictions for $t > t_0$; it will be predicted that (with high probability) $n(t)$ decreases as t increases, and that for long times $n(t) \simeq N/2$.

Now let the following data be given at $t = t_0$:

(ii) One second earlier, at $t = t_0 - 1$, all the particles were in the left-hand half. That is, $n(t_0 - 1) = N$.

Jaynes' principle, applied to this data, again yields irreversible predictions about future values of n; for $t > t_0$, $n(t)$ will be predicted (with high probability) to decrease toward $N/2$. (In fact, predictions for time t in case (ii) are identical with predictions for time $t + 1$ in case (i)).

Finally, let the data given at $t = t_0$ be:

(iii) One second in the future, at $t = t_0 + 1$, all the particles will be in the left-hand half. That is, $n(t_0 + 1) = N$.

Jaynes' principle applied to this data yields "reversed" predictions about n during t_0 to $t_0 + 1$. It will be predicted (with high probability) that $n(t)$ increases from $n(t_0) < N$ to $n(t_0 + 1) = N$ during t_0 to $t_0 + 1$.

The predictions in case (iii) are non-physical: we hardly ever observe $n(t)$ spontaneously increasing from $n(t_0) < N$ to $n(t_0 + 1) = N$. The non-physical predictions result from the non-physical nature of the assumed data: experimental information may refer to the present (as in case (i)) or the past (as in case (ii)), but never to the future. That is, it is impossible for an observer to obtain experimental data about the future; any assumed data which violates this principle leads to non-physical predictions. This asymmetry between past and future appears to be a fundamental feature of the physical world, not explainable in terms of any more fundamental principle.

Generalizing on the basis of this example, we have:

Non-Time-Symmetry of Classical Statistical Mechanics. From the facts that

(i) statistical mechanics deals with a probability distribution satisfying Liouville's equation;
(ii) the initial distribution is determined by Jaynes' principle;
(iii) experimental data never refers to the future;

it follows that there exist sequences of predictions which are consistent with statistical mechanics and for which the time-reversed sequence is *not* consistent. (5.2.11)

We will formulate this statement in a more specific and more quantitative way in Section 5.4.

Thus, predictions based on classical statistical mechanics are both relaxing and non-time-symmetric; hence classical statistical mechanics is irreversible.

The irreversibility of statistical *predictions* is not physically significant unless these predictions are likely to be experimentally *correct*. But we showed in Section 3.5 that, as a general rule, statistical predictions are highly likely to be correct provided that N is large. Thus, for large N, the experimentally observed behavior is highly likely to conform to the irreversible predictions. Briefly, many-body systems are highly likely to behave in the usual irreversible manner.

It is important to note that, in the preceding paragraph, "large N" does not mean "infinite N". That is, the statistical mechanical explanation of irreversibility does not rely on the assumption that N is infinite. It is clear, from simple physical examples, that any satisfactory statistical explanation of irreversibility must not assume $N \to \infty$. For instance, a set of pool balls behaves irreversibly (they are highly unlikely to spontaneously return to their original triangular arrangement) despite the fact that N is only 16. The effect of N on irreversibility will be brought out more clearly in Section 5.6, where we will give an exact analysis of a simple N-body system.

In order to gain further insight into the non-time-symmetry of statistical mechanics, consider once again the particle in a box. Let the measuring instruments be M_1, M_2, M_3, as described in Section 3.6. Recall that M_1 determines whether $0 < q < L/2$ or $L/2 < q < L$; M_2 determines whether $p > 0$ or $p < 0$; M_3 determines in which energy shell $nb < |p| < (n+1)b$ the particle lies. We will examine the evolution of $\langle q \rangle$ for several choices of the initial data. We will take the initial time as $t = 0$.

Case (i): At $t = 0$, the observer is given the information that at the present instant (i.e. at $t = 0$) the instruments show $0 < q < L/2$, $p > 0$, $0 < |p| < b$.

We have seen in Chapter 3 that Jaynes' principle then leads to the initial distribution of Figure 1, and that Liouville's equation implies that $\langle q \rangle_t$ evolves as in Figure 2.

Case (ii): At $t = 0$, the observer is given the information that at $4mL/b$ seconds earlier (i.e. at $t = -4mL/b$), the instruments showed $0 < q < L/2$, $p > 0$, $0 < |p| < b$.

It follows that at the initial instant ($t = 0$) the particle must be in the shaded region shown in Figure 3, so that Jaynes' principle implies that $\varrho(q, p, 0)$ is constant over this region and zero outside. Figure 4 shows the subsequent evolution of $\langle q \rangle_t$.

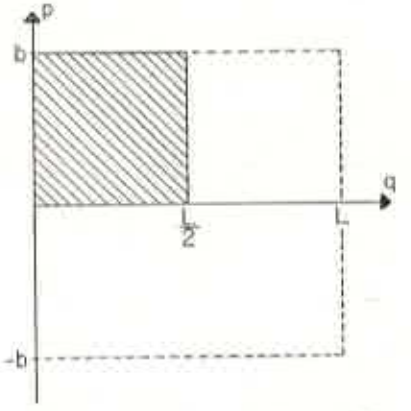

Figure 5.2-1 $\varrho(q, p, 0)$ for case (i)

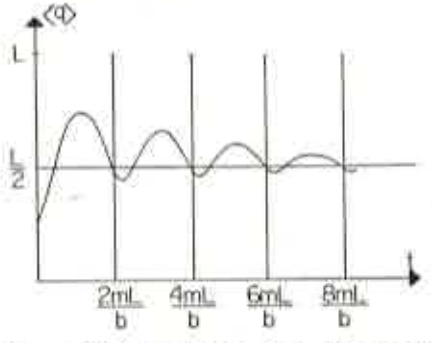

Figure 5.2-2 Evolution of $\langle q \rangle_t$ for case (i)

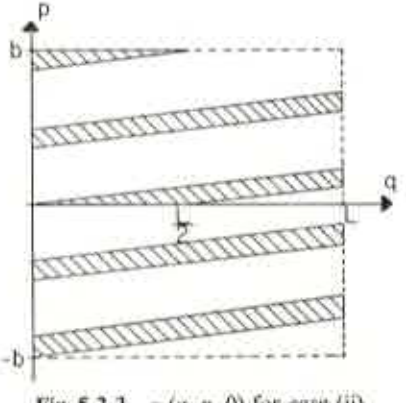

Fig. 5.2-3 $\varrho(q, p, 0)$ for case (ii)

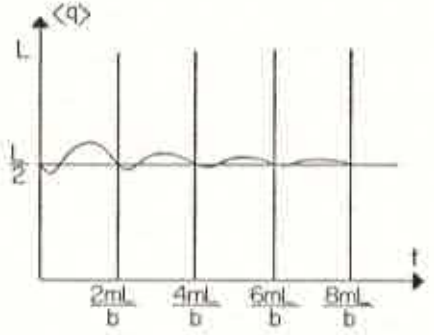

Figure 5.2-4 Evolution of $\langle q \rangle_t$ for case (ii)

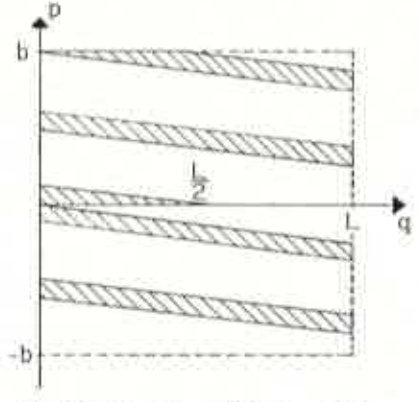

Figure 5.2-5 $\varrho\,(q, p, 0)$ for case (iii)

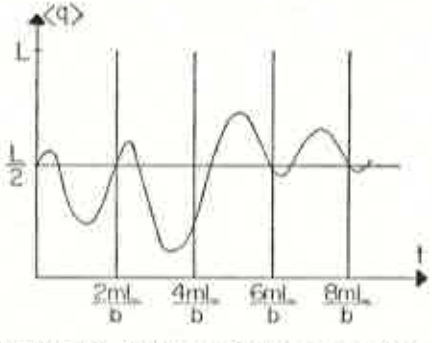

Figure 5.2-6 Evolution of $\langle q \rangle_t$ for case (iii)

Case (iii): At $t = 0$, we are given the information that at $4mL/b$ seconds *in the future* (i.e. at $t = 4mL/b$), the measuring instruments will show $0 < q < L/2$, $p > 0$, $0 < |p| < b$.

This data implies that at the initial instant ($t = 0$), the particle must be in the shaded region of Figure 5, so that Jaynes' principle implies that $\varrho\,(q, p, 0)$ is constant over this region. Figure 6 shows the subsequent evolution of $\langle q \rangle_t$.

In Figures 2 and 4 the oscillations of $\langle q \rangle_t$ die out with increasing t, but in Figure 6 the oscillations *increase* during $t = 0$ to $4mL/b$. Thus $\langle q \rangle_t$ continually approaches its equilibrium value in cases (i) and (ii), but departs from its equilibrium value during $t = 0$ to $t = 4mL/b$ in case (iii). The non-physical departure from equilibrium in case (iii) is due to the fact that the given information refers to the future rather than (as in cases (i) and (ii)) to the present or the past. Note that, even in case (iii), predictions must (according to the relaxation property of Liouville's equation) eventually relax to equilibrium. In case (iii), the relaxation toward equilibrium does not begin until $t = 4mL/b$.

It seems clear from this example that, if the given information refers to the present or the past, then statistical predications exhibit the usual approach to equilibrium, whereas if the information refers to a time t_0 seconds in the future, then predictions may exhibit a non-physical departure from equilibrium during $t = 0$ to t_0.

Summarizing the results of Sections 5.1 and 5.2: we can derive (i.e. explain) irreversibility on the basis of classical statistical mechanics, but not on the basis of pure mechanics. No special assumptions (such as $N \to \infty$, special initial conditions, special types of inter-particle interactions, or interactions with the environment) are needed in the derivation. All that is needed are the two basic principles of statistical mechanics (Jaynes' principle and Liouville's equation) plus the fact that experimental data never refers to the future.

5.3 IRREVERSIBILITY AND QUANTUM STATISTICAL MECHANICS

Quantum statistics is in many ways parallel to classical statistics, so we might expect that any discussion of quantum irreversibility would simply be a restatement, in quantum language, of classical irreversibility. As we will see, the situation is not that simple.

As regards time-symmetry, the quantum and classical situations are the same. From (5.1.2) we have (with $q \equiv (\mathbf{q}_1, \ldots, \mathbf{q}_N)$):

Time-Symmetry of the Quantum Liouville Equation. Let the initial mixture $\{\Psi_0^k(q)$ with probability $p_k\}$ evolve, during 0 to t, into $\{\Psi^k(q, t)$ with probability $p_k\}$. Then the initial mixture $\{\Psi^{k*}(q)$ with probability $p_k\}$ evolves, during 0 to t, into $\{\Psi^{k*}(q, -t)$ with probability $p_k\}$. (5.3.1)

Thus, if a certain sequence of predictions is consistent with the quantum Liouville equation, so is the time-reversed sequence.

Despite the time-symmetry of the quantum Liouville equation, it is clear that Jaynes' principle (when combined with the fact that experimental data never refers to the future) leads to the non-time-symmetry of quantum statistical mechanics. The argument here is precisely the same as the corresponding argument of Section 5.2. Thus, quantum statistical mechanics is non-time-symmetric.

We now ask: are quantum statistical predictions almost periodic? We have seen, in (5.2.9), that classical statistical predictions relax and hence cannot be almost periodic. On the other hand, in the quantum case we have:

Almost-Periodicity of Quantum Statistical Mechanics. Let the initial density operator have the form

$$\hat{\varrho}(0) = \sum_k |\Psi_0^k> p_k < \Psi_0^k|,$$

where the possible states $|\Psi_0^k\rangle$ form a discrete set, and where each possible state $|\Psi_0^k\rangle$ is expressible as a linear combination over a discrete set of energy eigenstates. Then $\hat{\varrho}(t)$ is almost periodic. That is, for any $\varepsilon > 0$, there exist infinitely many times τ (spread over the whole range $-\infty$ to ∞) such that $|\hat{\varrho}(\tau) - \hat{\varrho}(0)| < \varepsilon$. Here, the norm $|\hat{A}|$ of an operator is defined by $|\hat{A}|^2 \equiv \mathrm{Tr}\, \hat{A}^\dagger \hat{A}$. (5.3.2)

This result was first proved by Ono[3], and independently by Percival[9]. A proof is given in Appendix D. Thus, $\hat{\varrho}(t)$ keeps returning arbitrarily "near" to $\hat{\varrho}(0)$. It follows[3] (see Appendix D) that any quantum statistical expectation value $\langle A\rangle_t = \mathrm{Tr}\,(\hat{\varrho}\hat{A})$ keeps returning arbitrarily near to $\langle A\rangle_{t=0}$.

Thus there is a fundamental difference, as regards irreversibility, between classical and quantum statistics: classical predictions relax while quantum predictions are almost periodic.

As an example, consider an N-body gas in a box, and let the initial distribution correspond to the information that at $t = 0$, all N particles are in the left-hand half of the box. The initial distribution ($\varrho(x, 0)$ if the problem is

treated classically, $\hat{\varrho}(0)$ if the problem is treated quantum-mechanically) contains the information that "$n(0) = N$ with probability 1", where $n(t)$ means the number of particles in the left-hand half. Classically, $\varrho(x, t)$ evolves in such a way that, for all $t > \tau_r$ (τ_r = some "relaxation time"), the distribution $\varrho(x, t)$ contains the information "$|n(t) - N/2| < \varepsilon$ with probability $P \simeq 1$, where $\varepsilon \ll N$", i.e. it is nearly certain that $n(t) \simeq N/2$. But quantum-mechanically, $\hat{\varrho}(t)$ evolves in such a way that, at some "recurrence time" τ, all predictions are nearly equal to the initial predictions; hence $\hat{\varrho}(\tau)$ must contain the information "$|n(\tau) - N| < \varepsilon$ with probability $P \simeq 1$, where $\varepsilon \ll N$", i.e. it is nearly certain that at time τ all N particles will be back in the left-hand half. If the system is nearly classical (i.e. if classical statistical mechanics yields a good approximation to the "true" quantum situation), then we expect that $\tau \gg \tau_r$, and furthermore that τ will be very long on the lab scale.

Thus, there is a certain "coherence" in quantum statistical predictions: predictions might approach equilibrium over a finite period of time, but if we are just willing to project far enough into the future, we can re-obtain all the initial predictions. Note that the recurrence of quantum statistical predictions is a much stronger and experimentally more significant result than the simple mechanical recurrence of the classical phase point (the Poincaré recurrence theorem) or the quantum state vector. The mechanical recurrence theorem say merely that the initial mechanical state will recur. The quantum statistical recurrence theorem says that, not only will the initial quantum state recur (e.g. not only will all N particles eventually return to the left-hand half of the container), but that furthermore it is possible to *predict* with high probability, even on the basis of limited initial data, *when* the initial situation will recur (e.g. the time τ at which all particles will return to the left-hand half).

Thus, it appears that irreversibility is a general property of classical statistical mechanics but not of quantum statistical mechanics. More precisely, classical statistical mechanics exhibits relaxation, whereas quantum statistical mechanics does not. If we regard quantum mechanics as the true description of the physical world, then irreversibility holds approximately for all nearly-classical systems (i.e. systems whose true quantum behavior is approximated reasonably well by classical mechanics). For such systems, the time τ for predictable recurrences (or for any other predictable macroscopic deviation from the asymptotic equilibrium state) to occur must be very long on the lab scale. But for strongly non-classical systems, we have no reason to expect anything like irreversible behavior; the time τ for predictable macroscopic

fluctuations to occur could be of the order of the relaxation time, in which case nothing like relaxation to equilibrium will take place. Thus, irreversibility is not imbedded in the general theory of quantum statistics but *is* imbedded in the general theory of classical statistics. For strongly non-classical systems, the question "does the system exhibit nearly irreversible behavior?" can be answered only case-by-case, by looking at the statistical evolution of the particular system in question. *General* discussions of irreversibility must be carried out within the framework of classical statistics.

Now, most many-body systems are nearly classical, provided only that the temperature (more generally: the total energy) is not too low. Thus, irreversibility applies to falling waterfalls, boxes sliding down inclined planes, and the like. (More precisely: such systems exhibit irreversible behavior for times which are long on the laboratory time-scale. However, the quantum-statistical recurrence theorem implies that, even for nearly classical systems, quantum realities must eventually assert themselves and predictable recurrences must eventually take place if one is only willing to wait long enough. We expect that, for nearly classical many-body systems, the time required might be some large multiple of the age of the universe.) Such "few-body" systems as the 16 pool balls are also nearly classical, since the individual particles (the pool balls) are themselves nearly classical objects. Thus, the pool balls exhibit irreversibility; for example, it is highly unlikely that a scattered set of pool balls will spontaneously re-form into a triangular arrangement at rest with all the energy and momentum transferred to the cue ball. But a highly non-classical system (e.g. a many-atom system at low temperature, or any few-body system composed of atomic particles) might very well exhibit highly coherent, recurrent behavior.

Super-fluids and super-conductors are prime examples of many-body systems which, due to their highly quantum nature, do not exhibit irreversible behavior. These systems are structured in such a way that a macroscopic observer, possessing only limited control over (and hence limited information about) the initial state, can produce predictable recurrences in the state of the system. Such many-body recurrences can be produced in *classical* systems only if the observer has detailed control over the precise initial state. For example, in order to make the 16 pool balls return predictably to their initial triangular configuration with the cue ball absorbing all the energy, the observer would have to start the balls with *precisely* the right velocities. The significance of the quantum statistical recurrence theorem is that even a poorly informed observer can predict recurrences in a many-body system which is highly quantum mechanical.

Mathematically, there is an interesting formal similarity between classical and quantum systems as regards the question of almost-periodicity. The similarity is as follows:

(i) Both classical and quantum pure states are almost periodic.
(ii) Classical and quantum mixed states, where the mixture is discrete, are almost periodic.
(iii) Classical and quantum mixed states, where the mixture is *continuous*, appear to relax.

Result (i) is just a re-statement of the Poincaré and Ono recurrence theorems.[2,3] Result (ii) for quantum systems is just a re-statement of (5.3.2). In Appendix C, we give the proof of (ii) for the classical case.

Statement (iii) is only a conjecture, although a very reasonable one, for both classical and quantum systems. In Section 5.2, we showed that continuous mixtures relax for any classical system whose mechanical motion is unstable, and we argued that nearly every classical system of physical interest is unstable. Statement (iii) remains largely uninvestigated for quantum systems. Quantum statistical mixtures over a *continuous* set of states require a density operator formalism based on the notion of the eigendifferential,[10,11] and this makes matters rather complicated.

As an example of a finite (i.e. N finite, volume finite) quantum mechanical system described by a continuous mixture, consider a single one-dimensional particle in a box $0 \leqslant q \leqslant L$ (infinite potential well), and let the given data be $0 \leqslant q \leqslant L/2$. Jaynes' principle then implies that the initial mixture is spread out over all the coordinate eigenstates $|q\rangle$, where the eigenvalues q range over the continuum $0 \leqslant q \leqslant L/2$. This example is studied in Ref. 10, where it is tentatively concluded (but not proved) that the predictions relax. This type of example (finite system, described by a continuous mixture) is probably not of much practical importance, since in practice the given data always includes a statement about the energy (for instance, $\langle H \rangle$ might be given), and this will normally suffice to produce a *discrete* mixture.* For instance, in the above example of the particle in a box, it can be shown that if the given data is $0 \leqslant q \leqslant L/2$ and $\langle \hat{H} \rangle = E$ (E given), then Jaynes' principle implies that $\hat{\varrho}(0)$ is a mixture over a discrete set of states.

Thus, whether the system is classical or quantum mechanical, discrete mixtures are almost periodic while continuous mixtures relax. The fundamental difference between classical and quantum statistics (i.e. the relaxation of

* This fact was pointed to me by I.C. Percival in a private communication.

classical statistics versus the almost periodicity of quantum statistics) then resides solely in the fact that classical mixtures are generally continuous, whereas quantum mixtures are generally discrete. That is, we can trace the distinction back to the *continuous* nature of the classical world versus the *discrete* nature of the quantum world.

As a simple example, consider the spin 1/2 particle in a static external magnetic field B_0 directed along the z axis (see Section 4.2). Let the given data be

$$\langle s_x \rangle_{t=0} = A \quad (A \text{ given}). \tag{5.3.3}$$

Jaynes' principle leads to the initial density operator

$$\hat{\varrho}(0) = \frac{\exp(-\alpha \hat{s}_x)}{Z(\alpha)} \tag{5.3.4}$$

where, for proper normalization,

$$Z(\alpha) = 2\cosh(\alpha\hbar/2). \tag{5.3.5}$$

Since the Lagrange multiplier α must satisfy (3), it follows that

$$-(\hbar/2)\tanh(\alpha\hbar/2) = A. \tag{5.3.6}$$

From (4), it follows that $\hat{\varrho}(0)$ has the form of (4.3.8), with

$$\begin{aligned} p &= (\text{Prob. of } |+s_x\rangle) = \frac{\exp(-\alpha\hbar/2)}{Z(\alpha)} \\ q &= (\text{Prob. of } |-s_x\rangle) = \frac{\exp(+\alpha\hbar/2)}{Z(\alpha)} \end{aligned} \tag{5.3.7}$$

It follows from (7) and (4.3.11) that, in energy representation,

$$(\hat{\varrho}(t)) = \begin{pmatrix} 1/2 & -\tanh(\alpha\hbar/2)\exp(i\gamma B_0 t) \\ -\tanh(\alpha\hbar/2)\exp(-i\gamma B_0 t) & 1/2 \end{pmatrix} \tag{5.3.8}$$

where γ is the gyromagnetic ratio. Putting (7) into (4.3.12) and (4.3.13), we find that

$$\langle s_x \rangle_t = A\cos(\gamma B_0 t) \tag{5.3.9}$$

$$\sigma^2(s_x) = \hbar^2/4 - A^2\cos^2(\gamma B_0 t). \tag{5.3.10}$$

Equation (8) shows that the density operator is exactly periodic with frequency $\omega = \gamma B_0$; (9) shows explicitly that predictions about s_x are exactly periodic, and (10) shows that furthermore the reliability of those predictions is exactly periodic. (Note that (5.3.2) requires only that predictions be *almost* periodic; the exact periodicity obtained in this example is a special case of almost-periodicity).

5.4 THE GENERALIZED SECOND LAW

According to the preceding section, general discussions of the approach to equilibrium must be based on classical statistical mechanics. Thus, in the remainder of this chapter, we will consider only classical systems.

The second law of thermodynamics is an important generalization about the physical world. In its usual form, the second law is a statement only about equilibrium situations. It is obviously desirable to find a generalization of the second law which would apply also to non-equilibrium situations.

In order to find such a generalized second law, we must first extend the entropy concept to non-equilibrium situations. Thus, we seek a generalized entropy which reduces to the usual thermodynamic entropy (3.9.6) at thermal equilibrium. An obvious candidate for such a generalized, time-dependent entropy is

$$-k\int \varrho(x,t)\ln[h^{3N}N!\,\varrho(x,t)]\,dx.^{*} \tag{5.4.1}$$

Unfortunately, this expression is wrong. The problem is that (1) is time-independent, as we will now show: The distribution $\varrho(x,t)$ is a (time-dependent) integral of the motion. Hence the integrand of (1) is an integral of the motion. But the integrated value of any integral of the motion over any invariant region (i.e. any region which transforms into itself under the natural motion of the system) is constant. Since the entire phase space is an invariant region, (1) is constant.

Some authors[12,13] have assumed (1) to be the appropriate expression for the generalized entropy and have argued that, since (1) is time-independent for closed systems, it follows that only open systems exhibit irreversibility. But we have seen in Section 5.2 that, even for closed and finite systems, irreversibility is embedded in the structure of classical statistical mechanics. Thus, it should be possible to find a generalized entropy expression which is valid for closed systems.

* In case there are r different species of particles, $N!$ should be replaced by $N_1!\ldots N_r!$. If all the particles are distinguishable, the $N!$ should be omitted.

The most popular expression (see, for instance, Ref. 14) for a generalized entropy is

$$-k \int \varrho_{cg}(x,t) \ln [h^{3N} N! \varrho_{cg}(x,t)]\, dx. \tag{5.4.2}$$

Here, $\varrho_{cg}(x,t)$ is a "coarse-grained" probability distribution, defined in terms of $\varrho(x,t)$ by

$$\varrho_{cg}(x,t) \equiv \int_{R_i} \varrho(x',t)\, dx' \Big/ \int_{R_i} dx' \quad (x \in R_i), \tag{5.4.3}$$

where the phase regions or "cells" R_i form a partition of the phase space. Thus ϱ_{cg} is a step function taking on different values in different cells. It is possible in this way to obtain a generalized entropy which is time-dependent for closed systems. One drawback to this approach is that the coarse-grained entropy (2) is dependent on the manner in which the cells R_i are defined and there appears to be no physical reason for preferring one choice of the cells over another choice.[12,15] Another difficulty is that, if $\varrho(x,t)$ is canonical, then $\varrho_{cg}(x,t)$ is *not* canonical and hence (2) does not agree with the usual equilibrium entropy at thermal equilibrium. Thus, we will not adopt (2).

The properties which appear desirable in a generalized entropy $S(t)$ are:

(a) $S(t)$ reduces to the ordinary thermodynamic entropy at thermal equilibrium;

(b) $S(t)$ generally increases (although not necessarily monatonically—see Section 3.9) as the system approaches equilibrium;

(c) $S(t)$ is a "macroscopic observable", i.e. $S(t)$ depends only on the values of the observable quantities.

These properties say that $S(t)$ is actually a generalization of the equilibrium entropy, that $S(t)$ is a measure of the approach to or departure from equilibrium, and that $S(t)$ has experimental significance at the macroscopic level.

The following definition satisfies all three of the above criteria:

Generalized Entropy. Define $S(t)$ by

$$S(t) = -k \int \bar{\varrho}(x,t) \ln [h^{3N} N! \bar{\varrho}(x,t)]\, dx^*$$

where $\bar{\varrho}(x,t)$ is that distribution which maximizes the uncertainty subject to the predicted values of the observables at time t. That is, let $D(0)$ be the initial data concerning observables M, and let the corresponding initial distribution be $\varrho(x,0)$. After a time t, ϱ has evolved via Liouville's equation into $\varrho(x,t)$. Let the predictions about M at time t, on the basis of $\varrho(x,t)$, be $D(t)$. Then $\bar{\varrho}(x,t)$ maximizes the uncertainty (1) subject to $D(t)$. (5.4.4)

* See footnote, page 137.

This prescription satisfies criterion (a) above, since at thermal equilibrium the data is $\langle H(x) \rangle$, and maximization of the uncertainty subject to given $\langle H(x) \rangle$ yields the canonical distribution and hence (4) becomes the canonical entropy (3.9.6).

The definition satisfies (b) in the sense that, for any t,

$$S(t=0) \leqslant S(t) \leqslant S(t \to \infty), \tag{5.4.5}$$

and furthermore if the process is a non-equilibrium process then

$$S(t=0) < S(t \to \infty). \tag{5.4.6}$$

Proof of the first inequality of (5):[16] Expression (1), where $\varrho(x,t)$ obeys Liouville's equation, is time-independent. Since both $\varrho(x,0)$ and $\bar{\varrho}(x,0)$ maximize the uncertainty with respect to $D(0)$

$$\bar{\varrho}(x,0) = \varrho(x,0). \tag{5.4.7}$$

Both $\varrho(x,t)$ and $\bar{\varrho}(x,t)$ must be consistent with the predictions $D(t)$ but $\bar{\varrho}(x,t)$ maximizes (1) with respect to *all* distributions consistent with $D(t)$. Consequently,

$$\begin{aligned} S(t) &= -k \int \bar{\varrho}(t) \ln [h^{3N} N! \bar{\varrho}(t)]\, dx \\ &\geqslant -k \int \varrho(t) \ln [h^{3N} N! \varrho(t)]\, dx \\ &= -k \int \varrho(0) \ln [h^{3N} N! \varrho(0)]\, dx \\ &= -k \int \bar{\varrho}(0) \ln [h^{3N} N! \bar{\varrho}(0)]\, dx = S(0). \end{aligned}$$

Proof of the second inequality of (5): The predictions $D(t)$ must preserve the initial information concerning constants of the motion. If $\varrho(x,t)$ converges weakly to equilibrium as $t \to \infty$ (according to Section 5.2, this condition appears to hold for all physically interesting systems) then all data about "non-constants" of the motion is lost due to phase-mixing. Hence, $D(t)$ contains all the information contained in $D(t \to \infty)$. Since $\bar{\varrho}(x,t)$ maximizes (1) subject to $D(t)$ while $\bar{\varrho}(x, t \to \infty)$ maximizes (1) subject only to $D(t \to \infty)$, it follows that $S(t) \leqslant S(t \to \infty)$.

Proof of (6): For non-equilibrium processes, $D(0)$ must contain information about non-constants of the motion. Thus $D(0)$ contains all the informa-

tion in $D(t \to \infty)$, plus additional data about non-constants of the motion. Thus the inequality must hold.

The definition satisfies criterion (c), since once the macroscopic data $D(t)$ is given, $\bar{\varrho}(x, t)$ and hence $S(t)$ are determined.

For closed systems, equations (5) and (6) constitute the **generalized second law of thermodynamics**. (We will not consider open systems, in which "generalized heating" may be done on the system). Equation (5) says that there exists a function $S(t)$, dependent only on the predicted values of the observables, which takes on its minimum value at the instant $t = 0$ at which the initial data is obtained, and which attains its maximum value as $t \to \infty$. Equation (6) says that, unless the evolution of the system is a sequence of equilibrium states, the final value of S is necessarily greater than the initial value. Equation (5) is a direct generalization of the standard (equilibrium) form of the second law. The standard second law refers only to the initial and final predictions $D(0)$ and $D(t \to \infty)$. The generalized second law refers to the time-dependent predictions $D(t)$, and implies the standard second law for $D(0)$ and $D(t \to \infty)$.

The generalized second law implies that statistical mechanical predictions are non-time-symmetric. *Proof*: Assume that the following sequence of predictions $(A, B, C, \ldots)$ obeys (5), where $0 < t_1 < t_2$:

$$A \text{ at } t = 0, \quad B \text{ at } t_1, \quad C \text{ at } t_2, \ldots$$

Since this sequence obeys (5), we have (assuming that the inequality holds, rather than the equality) $S(A) < S(C)$, i.e. the entropy of the data C is greater than the entropy of the data A. Now consider the time-reversed sequence, beginning with data C:

$$C \text{ at } t = 0, \quad B \text{ at } t_1, \quad A \text{ at } t_2, \ldots.$$

Since $S(A) < S(C)$, this sequence does not obey (5). Hence this sequence is not permitted. This completes the general proof, promised in Section 5.2, that statistical mechanics is non-reflective.

The non-equilibrium entropy needn't be a monatonically increasing function of time. As was pointed out in Section 3.9, $S(t)$ may be non-monatonic and still agree with the standard form of the second law. In other words, during a portion of a non-equilibrium process the macroscopic predictions $D(t)$ could spontaneously move away from the asymptotic data $D(t \to \infty)$ without violating the standard second law. In Section 5.6, we will give an example illustrating the non-monatonic behavior of $S(t)$.

We will now examine the form of S for many-body systems. As was discussed in Section 3.5, the data for such systems ordinarily takes the form of expectation values. Thus the initial data has the form:

$$\langle g_i(x) \rangle_{t=0} = G_i(0) \quad (i = 1, 2, \ldots, m) \tag{5.4.8}$$

where the $G_i(0)$ are given. The macroscopic observables are then the $\langle g_i(x) \rangle$, and the predictions $D(t)$ are

$$G_i(t) = \langle g(x) \rangle_t = \int g_i(x)\, \varrho(x, t)\, dx \tag{5.4.9}$$

where $\varrho(x, t)$ is determined from $\varrho(x, 0)$ via Liouville's equation. According to (2.4.10), the distribution maximizing the uncertainty with respect to the data (8) is

$$\bar{\varrho}(x, 0) = Z^{-1}(0) \exp\left[-\sum_{i=1}^{m} \alpha_i(0)\, g_i(x)\right] \tag{5.4.10}$$

where

$$Z(0) = \int \exp\left[-\sum_{i=1}^{m} \alpha_i(0)\, g_i(x)\right] dx \equiv Z[\alpha_1(0), \ldots, \alpha_m(0)] \tag{5.4.11}$$

and the multipliers $\alpha_i(0)$ are determined in terms of the $G_i(0)$ by (8). At later times, the distribution $\bar{\varrho}$ maximizing the uncertainty with respect to the predictions (9) is

$$\bar{\varrho}(x, t) = Z^{-1}(t) \exp\left[-\sum_{i=1}^{m} \alpha_i(t)\, g_i(x)\right] \tag{5.4.12}$$

where

$$Z(t) = \int \exp\left[-\sum_{i=1}^{m} \alpha_i(t)\, g_i(x)\right] dx \equiv Z[\alpha_1(t), \ldots, \alpha_m(t)] \tag{5.4.13}$$

and where the multipliers $\alpha_i(t)$ are determined by the predictions (9), i.e. the $\alpha_i(t)$ are chosen to satisfy

$$G_i(t) = \int g_i(x)\, \bar{\varrho}(x, t)\, dx. \tag{5.4.14}$$

Thus, $\bar{\varrho}$ retains a "generalized canonical" form at all times (compare (3.6.6)), and the generalized entropy (4) becomes

$$S(t) = k \ln\left\{\frac{[Z[\alpha_1(t), \ldots, \alpha_m(t)]]}{h^{3N} N!}\right\} + \sum_{i=1}^{m} \alpha_i(t)\, G_i(t) \equiv S[G_1(t), \ldots, G_m(t)]. \tag{5.4.15}$$

Note that (by (7)) $\bar{\varrho}(x, 0) = \varrho(x, 0)$. However, at later times t we have $\bar{\varrho}(x, t) \neq \varrho(x, t)$, since $\varrho(x, t)$ does not retain a generalized canonical form

Several previous authors have used a generalized entropy either identical to or similar to the generalized canonical entropy defined by Eqs. (8)–(15).[16–22] This definition of the non-equilibrium entropy is implicit in the work of Jaynes.[16] The distribution $\bar{\varrho}(x, t)$ is the "simplest" distribution (i.e. the distribution containing the least information) which describes the predicted values of the observables. Thus, if we desire a theory which describes the evolution of the macroscopic observables and which, aside from this requirement, is as simple as possible, we should take $\bar{\varrho}(x, t)$ rather than $\varrho(x, t)$ as the fundamental distribution function of the theory. Using the initial condition (10), Robertson[22] has derived the equation of motion for $\bar{\varrho}(x, t)$; this procedure makes possible an efficient derivation, from Liouville's equation, of the exact equations of motion for the evolution of the macroscopic observables in a many-body system. The result is a closed set (i.e. the only unknowns are the macroscopic observables) of coupled, nonlinear, integrodifferential equations for the observables.[22]

5.5 MEANING OF THE GENERALIZED SECOND LAW

In this Section it will be necessary to distinguish statistical predictions from experimental outcomes. We will denote experimental data with an asterisk. For example, the predicted particle density at the point $\mathbf{q}$ at time t is $n(\mathbf{q}, t) = \langle n(\mathbf{q}; x)\rangle_t$ (see (3.4.8) and (3.4.9)). The experimentally measured density is then denoted $n^*(\mathbf{q}, t)$.

The generalized second law says that there exists a function of the predicted values of the observables, namely the function (5.4.4), which obeys the inequalities (5.4.5) and (5.4.6). This is a mathematical consequence of Liouville's equation and is true regardless of what interpretation we give to (5.4.4), (5.4.5) and (5.4.6). It is, however, interesting to interpret these equations.

The generalized entropy (5.4.4) has a very simple interpretation: *Letting M represent the measuring apparatus used in obtaining the initial experimental data $D^*(0)$, $S(t)$ is the observer's uncertainty about the phase point x at time t, when he knows only the instantaneous predictions $D(t)$ concerning M.* This interpretation follows from Jaynes' principle (Section 2.4); it is a direct generalization of the interpretation, in Section 3.9, of the ordinary thermodynamic entropy.

We usually establish $\varrho(x, 0)$ by altering a constraint in a system which was previously at equilibrium. Our interpretation of the standard second law (Section 3.9) said that, no matter how we alter the constraints at $t = 0$, the asymptotic predictions $D(t \to \infty)$ cannot contain more information (but

may contain less information) about x than did the initial data $D^*(0)$. The interpretation of the generalized second law is a straightforward generalization of this idea. The generalized second law states that *an observer whose measuring instruments M yield less than complete information can never gain information (but may lose information) about x by manipulating the constraints in a closed system; that is, the predictions $D(t)$ concerning the outcome of a measurement with M at time t contain less information about x than the experimental data $D^*(0)$ obtained with M at $t = 0$.*

The generalized second law (5.4.5) is an extremely broad principle. It applies to any mechanical system for which ϱ relaxes (in the sense of weak convergence—see Section 5.2) to equilibrium; even if ϱ does not relax, the first inequality of (5.4.5) is still valid and hence the interpretation of the generalized second law given in the preceding paragraph is still valid. The generalized second law broadens the standard second law in two important ways: the generalized second law applies to any set of observables (i.e. measuring instruments) M, rather than only to the standard thermodynamic variables; the generalized second law applies to every instant t during any non-equilibrium process, rather than only to the initial instant $t = 0$ and the asymptotic region $t \to \infty$.

Both the ordinary second law and the generalized second law are *statistical* laws; that is, they refer only to the *predicted* or *expected* values of the observables and not necessarily to the experimentally measured data. The ordinary second law is a relation between the initial thermodynamic data and the *expected* asymptotic values of the thermodynamic observables. It is obviously possible (although highly improbable if N is large) that a measurement made at some instant t (where $t \gg$ relaxation time) will yield data which deviates appreciably from the expected equilibrium values. Such deviations are called **statistical fluctuations**. Similarly, the generalized second law refers only to the expected values $D(t)$. A measurement at time t will yield an experimental result $D^*(t)$ which may or may not agree with the predictions $D(t)$; if N is large (larger than, say, 10—see Section 2.4), then it is highly probable that $D^*(t) = D(t)$ to within experimental accuracy. Thus, in any situation in which statistical mechanics yields statistically "sharp" (i.e. high probability) predictions, we may interpret the generalized second law in the following way in terms of the experimental data $D^*(t)$: *The experimental data $D^*(t)$, obtained with the measuring instruments M, will (with high probability) yield less information about the phase point x than did the initial experimental data $D^*(0)$ obtained with M.*

Broadly speaking, the generalized second law says that an observer can

lose information in many ways (by manipulation of the constraints), but can gain information only by making new measurements. In fact, according to the preceding paragraph, if the observer is restricted at time t to the same measuring apparatus M as was used for the initial measurement, then even additional measurements are highly unlikely to yield increased information; such measurements will (with high probability) agree with the predictions and hence cannot yield new information. In order to learn something new, it is necessary to bring in an improved measuring apparatus.

For example, if the apparatus M measures the number of particles n in the left-hand half of a box of gas, then the generalized second law says that the prediction $n(t)$ at time t contains less information about the precise mechanical state x than did the initial experimental data $n^*(0) = n(0)$. A measurement of n at time t will yield (with high probability) a value $n^*(t)$ which is nearly equal to the prediction $n(t)$, in which case $n^*(t)$ yields less information than did $n^*(0)$. In order to *increase* our information about x, an improved measuring instrument (such as a device capable of measuring the number of particles in a smaller portion of the box) must be brought in.

We can interpret the second law in terms of the orderliness of the system. In order to make this idea precise, we must give a precise definition of the concept of order. Rather than defining order, it is more useful to define disorder. We will speak of the "disorder of the mechanical state x with respect to the measuring instrument M", denoted $Q_M(x)$. We define $Q_M(x)$ by

$$Q_M(x) \equiv -k \int \varrho_x(x') \ln [h^{3N} N! \, \varrho_x(x')] \, dx',^* \tag{5.5.1}$$

where $\varrho_x(x')$ is that distribution corresponding (via Jaynes' principle) to the data which would be measured by M if the system were in the mechanical state x; note that x appears as a parameter in the distribution $\varrho_x(x')$.

For instance, if M measures the precise number of particles in the left-hand half of a box of gas, and if $n(x)$ represents the precise number of particles in the left-hand half when the phase point is x, then the disorder of the state x with respect to M is given by (1) where $\varrho_x(x')$ is that distribution maximizing $-\int \varrho(x') \ln [h^{3N} N! \, \varrho(x')] \, dx'$ with respect to all distributions satisfying the condition that the left-hand half of the box contains $n(x)$ particles. It is clear that, for this example, the minimum disorder is obtained for points x satisfying $n(x) = 0$ or N (all particles in either the left- or right-hand half of the box), and the disorder increases as $n(x)$ approaches $N/2$, reaching a maximum at $n(x) = N/2$. Note that the disorder $Q_M(x)$ is a *mechanical* concept

* See the footnote on page 137.

(i.e. it is a phase-function), and that it is relative to the measuring instrument. That is, a state x may be well ordered with respect to one measuring instrument but poorly ordered with respect to another. According to Jaynes' principle (Section 2.4), $Q_M(x)$ is the uncertainty which a macroscopic observer would have upon making a measurement with M, if the actual (but unknown) mechanical state was x. This seems to be a natural definition of the "disorder" concept.

With this definition of disorder, the generalized second law says: *If the initial data is obtained with the measuring instrument M, and if statistical mechanics yields statistically "sharp" predictions at time t, then it is highly likely that the mechanical state at time t will be less ordered with respect to M than was the initial mechanical state.* This statement is an immediate consequence of our previous interpretation of the generalized second law in terms of the experimental data $D^*(t)$. Thus the generalized second law implies that, for any measuring instrument M, and for any initial data $D^*(0)$, the mechanical state is likely to be less ordered (with respect to M) at time t than it was at $t = 0$. Note that this does not conflict with Poincaré's recurrence theorem, which says that eventually the mechanical system *will* come back to its initial state of "orderliness".

5.6 AN EXAMPLE: THE N-BODY IDEAL GAS[23]

The N-body ideal gas illustrates many of the concepts discussed in this book; it provides a transparent example of irreversibility in classical statistical mechanics. We can give an exact, non-equilibrium statistical mechanical analysis of this system for all t and for any N, and thus study in detail the approach to equilibrium.

Our system is N non-interacting point particles of mass m enclosed in a rectangular box with perfectly reflecting walls in zero external field. Physically, this highly idealized system is the zero-order approximation to a dilute, weakly interacting gas. Many important physical effects present in interacting systems will not show up in this simplified model; for example, the only relevant relaxation time for the ideal gas is the so-called "hydrodynamic relaxation time" (related to the length of the box), while in interacting systems the "kinetic relaxation time" (related to the range of the interaction) is also important. The main virtue of the ideal gas is that it is an exactly solvable system.

We will assume that the observables M consist of the expected values of the center of mass, total energy, and total momentum. These three quantities

are the only "global" quantities having much direct physical significance. Another possibility is the total angular momentum, but this quantity is difficult to deal with when the system is confined to a rectangular box. For simplicity, we will not consider such "local" quantities as the expected density at points $\mathbf{r}$ in the box.

Let the initial data be

$$\langle \mathbf{R}(x)\rangle_0,\ \langle H(x)\rangle_0,\ \langle \mathbf{P}(x)\rangle_0, \tag{5.6.1}$$

where

$$\mathbf{R}(x) \equiv N^{-1}\,\Sigma\,\mathbf{q}_j \tag{5.6.2}$$

$$H(x) \equiv (2m)^{-1}\,\Sigma\,\mathbf{p}_j\cdot\mathbf{p}_j \tag{5.6.3}$$

$$\mathbf{P}(x) \equiv \Sigma\,\mathbf{p}_j. \tag{5.6.4}$$

The phase space is defined by

$$0 \leqslant q_{j\mu} \leqslant L_\mu, \quad -\infty < p_{j\mu} < +\infty.$$

The index j ranges over the particles ($j = 1, 2, \ldots, N$), while μ ranges over the 3 directions ($\mu = 1, 2, 3$); L_μ is the length of the μth side of the box.

Jaynes' principle implies that the initial distribution is exponential in $\mathbf{R}$, H and $\mathbf{P}$ (compare (2.4.10)). Rewriting the Lagrange multipliers in a more convenient form, the initial distribution is

$$\varrho\,(x, 0) = \frac{1}{Z}\exp\left[-\boldsymbol{\alpha}_0\cdot\Sigma\,\mathbf{q}_j - \frac{\beta_0}{2m}\,\Sigma\,(\mathbf{p}_j - \mathbf{b}_0)^2\right], \tag{5.6.5}$$

where $\boldsymbol{\alpha}_0$, β_0 and $\mathbf{b}_0$ are chosen to satisfy the data (1), and where by direct integration

$$Z = \int \exp\left[-\boldsymbol{\alpha}_0\cdot\Sigma\,\mathbf{q}_j - \frac{\beta_0}{2m}\,\Sigma\,(\mathbf{p}_j - \mathbf{b}_0)^2\right] dx = \left(\prod_{\mu=1}^{3} Z_{1\mu}\right)^N, \tag{5.6.6}$$

$$Z_{1\mu} = \left(\frac{2\pi m}{\beta_0}\right)^{1/2}\frac{1 - \exp(-\alpha_{0\mu}L_\mu)}{\alpha_{0\mu}}. \tag{5.6.7}$$

The parameters β_0 and $\mathbf{b}_0$ have simple interpretations: by direct integration,

$$\langle \mathbf{p}_j\rangle_{t=0} = \mathbf{b}_0, \tag{5.6.8}$$

$$\left\langle \frac{(\mathbf{p}_j - \mathbf{b}_0)^2}{2m}\right\rangle_{t=0} = \frac{3}{2\beta_0}, \tag{5.6.9}$$

so that $\mathbf{b}_0$ is the expected momentum of any particle at $t = 0$, and $3/2\beta_0$ is the expected kinetic energy per particle at $t = 0$ as seen by an observer moving at the mean velocity $\mathbf{b}_0/m$, i.e. $3/2\beta_0$ is the expectation value of the "random kinetic energy". If a non-equilibrium temperature $T(t)$ is defined by the statement that, for point particles, the expected random kinetic energy per particle is $(3/2)\,kT(t)$, then $\beta_0 = 1/kT(0)$.

The initial distribution (5) would be an appropriate description of a burst of non-interacting particles, shot into a box (initially open at one end, and then closed at $t = 0$), where the total energy is known, and where the position and velocity of the center of mass at $t = 0$ are known.

By (5), the particles are statistically independent at $t = 0$; by (3), they are dynamically independent. It follows (see Section 3.4) that the particles are statistically independent for all t. Hence, the N-body problem reduces to N one-body problems; this is, of course, the simplifying feature which makes the problem exactly solvable.

Using a technique originally suggested by Born,[24] we can find the exact, analytic solution to Liouville's equation for the initial condition (5). We won't go through the details of calculation here, but will just give the result (the details may be found in Ref. 23). The result is

$$\varrho\,(x, t) = \prod_{j=1}^{N}\prod_{\mu=1}^{3}\varrho^{(1)}(q_{j\mu}, p_{j\mu}, t), \tag{5.6.10}$$

where the reduced distribution $\varrho^{(1)}$ is given by

$$\begin{aligned}\varrho^{(1)}(q_{j\mu}, p_{j\mu}, t) = {}& \frac{s_0(p_{j\mu})}{Z_{1\mu}}\,\frac{1 - \exp(-\alpha_{0\mu}L_\mu)}{2\alpha_{0\mu}L_\mu} \\ & + \frac{s_0(p_{j\mu})}{Z_{1\mu}}\sum_{n=1}^{\infty}\frac{\alpha_{0\mu}L_\mu}{\alpha_{0\mu}^2L_\mu^2 + n^2\pi^2}\,[1 - (-1)^n\exp(-\alpha_{0\mu}L_\mu)]\times \\ & \times\cos\left[\frac{n\pi}{L_\mu}\left(q_{j\mu} - \frac{p_{j\mu}t}{m}\right)\right] + \frac{s_1(p_{j\mu})}{Z_{1\mu}}\sum_{n=1}^{\infty}\frac{n\pi}{\alpha_{0\mu}^2L_\mu^2 + n^2\pi^2}\times \\ & \times[1 - (-1)^n\exp(-\alpha_{0\mu}L_\mu)]\sin\left[\frac{n\pi}{L_\mu}\left(q_{j\mu} - \frac{p_{j\mu}t}{m}\right)\right]\end{aligned} \tag{5.6.11}$$

where $s_0(p_{j\mu})$ and $s_1(p_{j\mu})$ are defined by

$$s_k(p_{j\mu}) = \exp\left[\frac{-\beta_0}{2m}\,(p_{j\mu} - b_{0\mu})^2\right] + (-1)^k\exp\left[\frac{-\beta_0}{2m}\,(p_{j\mu} + b_{0\mu})^2\right], \quad (k = 0, 1). \tag{5.6.12}$$

According to (10) and (11), $\varrho(x, t)$ is an oscillatory function of time at any fixed phase point x; $\varrho(x, t)$ does not possess a limit as $t \to \infty$, i.e. ϱ does not converge strongly to any asymptotic distribution as $t \to \infty$ (see Section 5.2). However, phase-mixing (see Section 5.2) does occur, and the distribution converges weakly to equilibrium. We can show this explicitly by calculating the moments $\langle (q_{j\mu})^r (p_{j\mu})^s \rangle$ $(r, s = 1, 2, 3, \ldots)$. By direct integration, using the reduced distribution (11), we find[21]

$$\langle (q_{j\mu})^r (p_{j\mu})^s \rangle_t = \langle (q_{j\mu})^r (p_{j\mu})^s \rangle_\infty + F_{rs}(t) \tag{5.6.13}$$

where the asymptotic expectation values are given by

$$\langle (q_{j\mu})^r (p_{j\mu})^s \rangle_\infty \equiv \int (q_{j\mu})^r (p_{j\mu})^s \varrho_{eq}^{(1)}(q_{j\mu}, p_{j\mu})\, dq_{j\mu}\, dp_{j\mu}, \tag{5.6.14}$$

$$\varrho_{eq}^{(1)}(q_{j\mu}, p_{j\mu}) = \left(\frac{\beta_0}{2\pi m}\right)^{1/2} \frac{s_0(p_{j\mu})}{2L_\mu} \tag{5.6.15}$$

and where the time-dependent part $F_{rs}(t)$ goes to zero as $t \to \infty$. For $t > \sqrt{2\beta_0 m}\, L_\mu/\pi$ the dominant term in $F_s(t)$ has the form

$$F_{rs}(t) \simeq (\text{const.})\, t^s \exp\left(\frac{-\pi^2 t^2}{2\beta_0 L_\mu^2 m}\right). \tag{5.6.16}$$

(The exact expression for $F_s(t)$ is complicated and will be written out here.) It follows from (13), (14) and (15) that, for any phase-function $g(x)$ expressible as a uniformly convergent Taylor expansion in the $q_{j\mu}$ and $p_{j\mu}$, the expectation value relaxes to equilibrium:

$$\langle g(x) \rangle_t \to \langle g(x) \rangle_{eq} \equiv \int g(x)\, \varrho_{eq}(x)\, dx, \tag{5.6.17}$$

$$\varrho_{eq}(x) = \prod_{j=1}^{N} \prod_{\mu=1}^{3} \varrho_{eq}^{(1)}(q_{j\mu}, p_{j\mu}). \tag{5.6.18}$$

The distribution defined by (12), (15) and (18) is an equilibrium distribution, i.e. a distribution which is a function only of the constants of the motion $(p_{j\mu})^2$. Thus, the exact distribution $\varrho(x, t)$ converges weakly, but not strongly, to equilibrium. All expectation values settle down to the equilibrium values given by $\varrho_{eq}(x)$.

By (13) and (16), the coordinate moments $\langle (q_{j\mu})^r \rangle_t$ all relax as

$$\exp(-\pi^2 t^2 / 2\beta_0 L_\mu^2 m);$$

this function has a relaxation time

$$\tau_\mu = \sqrt{2m\beta_0}\, \frac{L_\mu}{\pi} \quad \text{(relaxation time for } \langle (q_{j\mu})^r \rangle\text{)}. \tag{5.6.19}$$

The momentum moment $\langle (p_{j\mu})^s \rangle_t$ has a relaxation time

$$(\sqrt{2} + \sqrt{s})\, \sqrt{m\beta_0}\, \frac{L_\mu}{\pi} \quad \text{(relaxation time for } \langle (p_{j\mu})^s \rangle\text{)} \tag{5.6.20}$$

since (16) has a maximum at $\sqrt{sm\beta_0}\, L_\mu/\pi$ and then falls off as $\exp(-at^2)$ with the relaxation time $\sqrt{2m\beta_0}\, L_\mu/\pi$. Thus, higher momentum moments fall off more slowly, so that phase functions having a simple momentum dependence (i.e. whose Taylor expansions contain only low powers of the momenta) relax rapidly, whereas phase functions having a complicated momentum dependence relax slowly. The reason for this is that Gibbs phase mixing implies a stretching out of the initial distribution into finer and finer filaments, where the filaments (for particles in a perfectly reflecting box) are nearly parallel to the $q_{j\mu}$ axis (see Figure 3.3-2). The condition for an expectation value $\langle g(x) \rangle$ to be relaxed is that $g(x)$ must vary only a small amount as x varies from one filament to another. Thus, $\langle g(x) \rangle$ will relax slowly if $g(x)$ varies rapidly with $p_{j\mu}$, i.e. if $g(x)$ has a complicated momentum dependence.

As will now be shown, the relaxation time (19) is determined by the length L_μ of the box and the initial dispersion σ_μ of the one-particle velocity. Using (8), (9), and (19),

$$\sigma_\mu^2 \equiv \left\langle \left(\frac{p_\mu}{m} - \left\langle \frac{p_\mu}{m} \right\rangle \right)^2 \right\rangle_{t=0} = \frac{1}{m^2} \langle (p_\mu - b_{0\mu})^2 \rangle_{t=0}$$

$$= \frac{1}{\beta_0 m} = \frac{2L_\mu^2}{\pi^2 \tau_\mu^2},$$

so that

$$\tau_\mu = \frac{\sqrt{2}\, L_\mu}{\pi \sigma_\mu}. \tag{5.6.21}$$

Thus (neglecting the factor $\sqrt{2}/\pi$) τ_μ is the time required for a particle moving at speed σ_μ to cross the box in the μ direction. As seen in the center of mass system, τ_μ is the time required for the particles to spread out (due to their "random velocities") over a distance L_μ.

By direct integration, using the exact distribution function given by (10), (11), (12), we find the following exact expressions for the evolution of the

observables $\langle \mathbf{R} \rangle$, $\langle H \rangle$, $\langle \mathbf{P} \rangle$:

$$\langle R_\mu \rangle_t = \langle q_{1\mu} \rangle_t, \tag{5.6.22}$$

$$\langle H \rangle_t = \langle H \rangle_{t=0}, \tag{5.6.23}$$

$$\langle P_\mu \rangle_t = N \langle p_{1\mu} \rangle_t, \tag{5.6.24}$$

where the expected one-particle position and momentum are given by

$$\langle q_1 \rangle_t = \frac{L}{2} - \frac{4\alpha_0 L^2}{\pi^2 \tanh(\alpha_0 L/2)} \sum_{n \text{ odd}} \frac{1}{n^2 (\alpha_0^2 L^2 + n^2\pi^2)^{1/2}} \times \times \exp\left(\frac{-n^2 t^2}{\tau^2}\right) \cos\left(\frac{n\pi b_0 t}{Lm} + \Phi_n\right), \tag{5.6.25}$$

$$\langle p_1 \rangle_t = \frac{4\alpha_0 L}{\tanh(\alpha_0 L/2)} \sum_{n \text{ odd}} \frac{1}{\alpha_0^2 L^2 + n^2\pi^2} \exp\left(\frac{-n^2 t^2}{\tau^2}\right) \times \times \left[\sin\left(\frac{n\pi b_0 t}{Lm}\right)\left(\frac{n\pi t}{\beta_0 L} - \frac{\alpha L b_0}{n\pi}\right) + \cos\left(\frac{n\pi b_0 t}{Lm}\right)\left(\frac{\alpha_0 t}{\beta_0} + b_0\right)\right]. \tag{5.6.26}$$

For simplicity, the subscript μ is omitted in (25) and (26); the phase angles Φ_n in (25) are given by

$$\cos \Phi_n = \frac{\alpha L}{(\alpha^2 L^2 + n^2\pi^2)^{1/2}}.$$

Thus, $\langle R_\mu \rangle$ and $\langle P_\mu \rangle$ both undergo a damped oscillatory approach to equilibrium. According to (25) and (26), the period of the oscillations is $2L_\mu m/b_{0\mu}$ (this is the time required for a particle moving at the mean velocity $b_{0\mu}/m$ to cross the box twice in the μ direction) and the relaxation time is τ_μ given by (19) or (21). If the dispersion σ_μ in velocities is small compared to the mean velocity $b_{0\mu}/m$ (i.e. if the initial burst of particles has a well-defined velocity $\mathbf{b}_0/m$) then the particles bounce back and forth in the box many times before settling down to equilibrium. As the system settles down to equilibrium, the initial ordered kinetic energy turns into disordered kinetic energy, not observable as kinetic energy by the observer (who is limited to observing $\langle \mathbf{R} \rangle$, $\langle H \rangle$ and $\langle \mathbf{P} \rangle$).

We will now investigate the generalized entropy. Since the data is given in terms of expectation values, the formalism of (5.4.8)–(5.4.15) is applicable. The distribution $\bar{\varrho}(x, t)$ (see (5.4.12)) has the same form as (5):

$$\bar{\varrho}(x, t) = Z^{-1}(t) \exp\left[-\alpha(t) \cdot \Sigma \mathbf{q}_j - \frac{\beta(t)}{2m} \Sigma (\mathbf{p}_j - \mathbf{b}(t))^2\right]. \tag{5.6.27}$$

The coefficients $\alpha(t)$, $\beta(t)$, $\mathbf{b}(t)$ are determined by the condition that $\bar{\varrho}(x, t)$ must yield the predicted values (22)–(24) of the observables. Thus (compare (8), (9))

$$\mathbf{b}(t) = \langle \mathbf{p}_j \rangle_t, \tag{5.6.28}$$

$$\frac{3}{2\beta(t)} = \left\langle \frac{(\mathbf{p}_j - \mathbf{b}(t))^2}{2m} \right\rangle_t. \tag{5.6.29}$$

If the non-equilibrium temperature $T(t)$ is defined by the statement that $3kT(t)/2$ is the expected kinetic energy per particle as seen from the center of mass frame, then (29) implies that $\beta(t) = 1/kT(t)$.

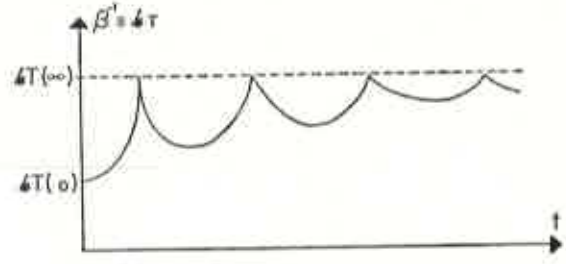

Figure 5.6-1 Time dependence of $\beta^{-1}(t) \equiv kT(t)$. The asymptotic value is given by (5.6.35)

It may be shown[23] that the coefficients $\alpha(t)$, $\beta(t)$, $\mathbf{b}(t)$ are related directly to the predictions (22)–(24) by

$$\mathbf{b} = \frac{\langle \mathbf{P} \rangle}{N} \tag{5.6.30}$$

$$\frac{3}{2\beta} + \frac{\mathbf{b} \cdot \mathbf{b}}{2m} = \frac{\langle H \rangle}{N}, \tag{5.6.31}$$

$$\frac{1}{\alpha_\mu} - \frac{L_\mu}{\exp(\alpha_\mu L_\mu) - 1} = \langle R_\mu \rangle. \tag{5.6.32}$$

By (5.4.7) we have $\alpha_0 = \alpha(t = 0)$, $\beta_0 = \beta(t = 0)$, $\mathbf{b}_0 = \mathbf{b}(t = 0)$. Using (30), (31), and the results (22)–(26), it may be seen that $\beta^{-1}(t) \equiv kT(t)$ has the behavior shown in Figure 1 (drawn for the special case of a cube-shaped container $L_1 = L_2 = L_3$). Thus, the non-equilibrium temperature increases (although not monotonically), due to the conversion of ordered kinetic energy into disordered kinetic energy. (If the 3 sides are not of equal length, $\beta^{-1}(t)$ will increase somewhat more smoothly than in Figure 1).

Using (27), the generalized entropy (5.4.4) is given in terms of the coefficients $\alpha(t), \beta(t), \mathbf{b}(t)$ by

$$S(t) = \frac{3}{2} Nk \ln\left(\frac{2\pi m}{\beta} V^{2/3}\right) + \frac{3}{2} Nk - k \ln(h^{3N} N!) + Nk \sum_{\mu=1}^{3} \left\{ \ln\left[\frac{1 - \exp(-\alpha_\mu L_\mu)}{\alpha_\mu L_\mu}\right] + 1 - \frac{\alpha_\mu L_\mu}{\exp(\alpha_\mu L_\mu) - 1} \right\} \tag{5.6.33}$$

where $V = L_1 L_2 L_3$ is the volume of the box. As $t \to \infty$, $\langle R_\mu \rangle_t \to L_\mu/2$ (by (22) and (25)) and hence (by (32)) $\alpha_\mu(t) \to 0$. But the last term of (33) vanishes as $\alpha_\mu \to 0$, and hence

$$\lim_{t\to\infty} S(t) = \frac{3}{2} Nk \left[\frac{2\pi m V^{2/3}}{\beta(\infty)}\right] + \frac{3}{2} Nk - k \ln(h^{3N} N!) \tag{5.6.34}$$

where (by (26), (30) and (31))

$$\frac{1}{\beta(\infty)} = \frac{2\langle H \rangle}{3N}. \tag{5.6.35}$$

Equations (34) and (35) are the conventional expressions for the entropy and kT for an ideal gas at thermal equilibrium.

It may be shown[23] that the non-equilibrium entropy evolves as in Figure 2, drawn for the special case that $L_1 = L_2 = L_3$. (If the three sides are not of equal length, then the non-monatonic increase shown in Figure 2 will be somewhat smoother). Figure 2 illustrates the generalized second law

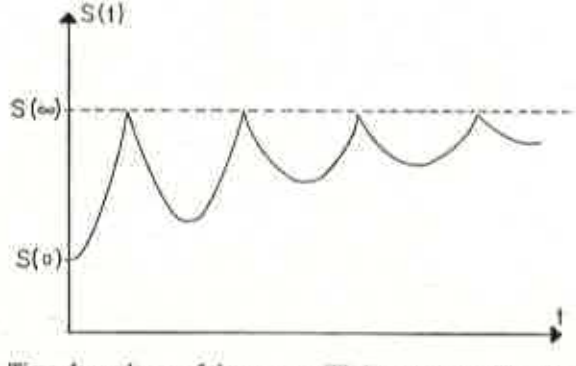

Figure 5.6-2 Time-dependence of the non-equilibrium entropy. The asymptotic value is given by (5.6.34)

(5.4.5) and (5.4.6), and also shows that $S(t)$ does not necessarily increase monatonically. In the present example, oscillations of S are due to reflections at the wall: as the N particles are reflected at a wall, the momenta are at first rapidly disordered (since some particles have reflected and others have not) and then rapidly reordered (as the remaining particles are reflected). Note that the oscillations of S occur on a time scale which is short compared to the relaxation time of the system; on the longer time-scale, the trend of $S(t)$ is upward.

Our interpretation of the generalized second law (see Section 5.5) says that there is less information about x in the predictions $\langle \mathbf{R} \rangle_t$, $\langle H \rangle_t$, $\langle \mathbf{P} \rangle_t$ than there was in the initial data $\langle \mathbf{R} \rangle_0$, $\langle H \rangle_0$, $\langle \mathbf{P} \rangle_0$; furthermore, there is less information in the asymptotic predictions

$$\langle \mathbf{R} \rangle_\infty = L/2, \quad \langle H \rangle_\infty = \langle H \rangle_0, \quad \langle \mathbf{P} \rangle_\infty = 0,$$

than there was in $\langle \mathbf{R} \rangle_t$, $\langle H \rangle_t$, $\langle \mathbf{P} \rangle_t$. If N is reasonably large, so that the predictions are highly likely to agree with experiment, then it is highly likely that the experimental data $\mathbf{R}$, H, $\mathbf{P}$, will yield less information at time t than at $t = 0$, and less information at $t = \infty$ (i.e. at any $t \gg \tau_\mu$) than at finite times (i.e. at times $t < \tau_\mu$).

The present example illustrates the fact that, although closed finite systems generally relax (in the sense that the distribution function converges weakly) to equilibrium, the asymptotic distribution is not necessarily canonical; that is, the system need not relax to *thermal* equilibrium. In fact, the asymptotic distribution given by (12), (15), and (18) is an equilibrium distribution but it is not canonical; it would be canonical only if the initial expected momentum were zero, i.e. only if $\mathbf{b}_0 = 0$; otherwise, the function $s_0(p_{j\mu})$ is peaked around $\pm b_{0\mu}$. This failure to achieve thermal equilibrium is partially due to our neglect of interactions between particles, and partially due to our neglect of interactions with the external world. The Hamiltonian (3) possesses $3N$ constants of the motion, namely the $(p_{j\mu})^2$; the distribution function $\varrho(x, t)$ never "forgets" the distribution of these constants of the motion, and hence the asymptotic distribution must retain all the initial information about the $(p_{j\mu})^2$. The introduction of a small interaction between the particles (for instance, a small but finite particle radius, so that particles could collide) should destroy these constants of the motion so that the only remaining constant of the motion would then be $H(x)$. The asymptotic distribution (to which $\varrho(x, t)$ is weakly convergent) would then depend only on $H(x)$, and would retain the initial information about $H(x)$. But this distribution would still be non-canonical, since according to (5) the energy is not canonically distributed

initially. The introduction of an appropriate interaction with the external world would then, presumably, rearrange the energy distribution in such a way that ϱ would converge weakly to the canonical distribution.

If the interactions between particles, and the interaction with the external world, are sufficiently small, then the relaxation to a final canonical distribution will occur over a time which is long compared to the time τ_μ (given by (19)) for the system to relax to the non-canonical distribution given by (15) and (18). In this case, the system will first relax to $\varrho_{eq}(x)$ given by (15) and (18) and then $\varrho_{eq}(x)$ will gradually relax to the canonical distribution.

All of the results in this Section are valid for arbitrary N. In particular, ϱ converges weakly, and all predictions relax to equilibrium, whether N is 1, or 10^{23}, or ∞. However, the predictions are likely to be *correct* (i.e. to agree with measured values) only if N is reasonably large. For example, by (22) the predicted center of mass $\langle R_\mu \rangle_t$ is entirely independent of N. However, the statistical variance in R_μ is

$$\langle (R_\mu - \langle R_\mu \rangle)^2 \rangle_t = \frac{1}{N} \langle (q_{1\mu} - \langle q_{1\mu} \rangle)^2 \rangle_t \tag{5.6.36}$$

and is highly dependent on N. The dispersion (square root of the variance) of R_μ is $1/\sqrt{N}$ times the dispersion of $q_{1\mu}$. For $t \gg \tau_\mu$, the distribution of $q_{1\mu}$ is approximately constant over $0 < q_{1\mu} < L_\mu$, so that the dispersion of $q_{1\mu}$ is approximately $L_\mu/\sqrt{12}$. Hence, for $t \gg \tau_\mu$, the dispersion of R_μ is $L_\mu/\sqrt{12N}$. According to the Tchebycheff inequality (2.1.45), we can state with high confidence that R_μ is within a few dispersion distances of $\langle R_\mu \rangle$. Thus, for $N = 100$ and $t \gg \tau_\mu$, we can predict with high confidence that $0.4L_\mu < R_\mu < 0.6L_\mu$. If $N = 10{,}000$ and $t \gg \tau_\mu$, we can predict with high confidence that $0.49L_\mu < R_\mu < 0.51L_\mu$.

5.7 IRREVERSIBILITY, COSMOLOGY, AND TIME

We will offer a few observations about the significance of irreversibility in cosmology and in the concept of time. There exists an immense literature on these questions; for a review of the literature before 1965, see Ref. 25. Our discussion will be quite brief and restricted; we look only at those aspects of cosmology and time which are directly related to the ideas developed in this book.

In order to keep the discussion reasonably simple, assume the universe to be a closed, classical mechanical system. (This assumption is open to question, naturally). What does the second law of thermodynamics imply about this system?

The first observation is that the standard second law implies nothing about this system, since the standard second law refers only to thermal equilibrium and it would obviously be inappropriate to attempt to describe the universe by a thermal equilibrium (i.e. canonical) distribution.

Thus, we must turn to the generalized second law. But the generalized second law has no content until the observables M have been specified. For the sake of continuing the argument, suppose that the observables are the mass densities throughout some finite region R of 3-dimensional space (R might represent that part of the universe which is observable by a ground-based astronomer). We will ignore problems introduced by the finite speed of light, and assume that the observer is able to make an instantaneous macroscopic observation of the phase function $f(\mathbf{q}; x)$ ($\mathbf{q} \in R$) representing the mass density at $\mathbf{q}$ when the mechanical phase point of the universe is x. Thus the observed data has the form

$$\langle f(\mathbf{q}; x) \rangle_{t=0} = F(\mathbf{q}), \tag{5.7.1}$$

where $F(\mathbf{q})$ is known for all $\mathbf{q}$ contained in R.

The generalized second law says that the "disorder of the universe with respect to the observable density" is highly likely (on the basis of data measured at $t = 0$) to be greater at a later time t_1 than it was at $t = 0$. That is, for an observer who is able to measure the macroscopic density throughout R, it is highly likely (on the basis of a measurement at $t = 0$) that the density will be more uniform (i.e. nearer to a constant through R) at later times. Thus it is highly likely, on the basis of presently known data, that the density throughout R will be more uniform in the future than it is now. This does not contradict the Poincaré recurrence theorem, which asserts that eventually the universe will return to a mechanical state approximately equal to its present state. On the basis of any reasonable presently measured data about the mass density it is highly likely (although not certain!) that the recurrence time will be very long.

Thus, on the basis of known data, the universe is highly likely to be more "run down" (i.e. have a more uniform density) in the future than at present. But note that this is only a probabilistic statement. There may be circumstances which are relevant to the future behavior of the universe but which do not appear in the known data, and which imply that tomorrow morning the universe will begin evolving toward a more ordered state with respect to

the density in the region R (see Section 3.7). Furthermore, it is certain (according to Poincaré's recurrence theorem) that this evolution toward a more ordered state *will* eventually occur. Our measuring instruments aren't sufficiently sensitive to pick up information about *when* this will occur, so our statements about the *expected* future behavior of the universe cannot contain predictions about an eventual evolution toward a more ordered state.

Concerning the concept of time, it is clear that the generalized second law is related in some manner to the question of the "direction" of time. It is sometimes asserted that the second law explains the distinction between past and future, or that the future may be defined as the direction of increasing entropy. This assertion says that the second law is more fundamental than the distinction between past and future.

It seems to the author that the above assertion is wrong. It was seen in Section 5.2 that the generalized second law is derived *from* the distinction between past and future; hence the distinction between past and future is more fundamental than the second law.

The following statement seems to be the most fundamental physical assertion which can be made regarding past and future: we can classify all instants t into two categories; the first category contains those instants about which experimental data is (or could be) known, and the second category contains those instants about which no experimental data is known. The first category is conventionally called the "past" and the second is called the "future". The instants may be labeled with real numbers running from $-\infty$ to $+\infty$, in such a way that the "past" instants constitute a set of the form $(-\infty < t < t_0)$, and the "future" instants constitute the set $(t_0 < t < +\infty)$. The choice of the positive direction as the future is purely a convention.

According to Sections 5.2 and 5.4, irreversibility and the generalized second law are derivable from the existence of the above two categories of instants: an "information-gathering category" (the past), and a "predictive category" (the future). The existence of these two categories seems to be a fundamental feature of nature, not explainable in terms of the second law or in terms of any other physical law.

REFERENCES

1. H. Bohr, *Almost Periodic Functions*, Chelsea Publishing Company, New York, 1947.
2. G. E. Uhlenbeck and G. W. Ford, *Lectures in Statistical Mechanics* (American Mathematical Society, Providence, R.I., 1963).
3. S. Ono, *Mem. Fac. Eng. Kyushu Univ.* **11**, 125 (1949).
4. P. Bocchieri and A. Loinger, *Phys. Rev.* **107**, 337 (1957).

5. F. Riesz and B. Sz.-Nagy, *Functional Analysis*, (Ungar Publishing Company, New York, 1955).
6. E. W. Hobson, *The Theory of Functions of a Real Variable* (Cambridge University Press, Cambridge, 1926).
7. A. Wintner, *The Analytical Foundations of Celestial Mechanics* (Princeton University Press, Princeton, New Jersey, 1941).
8. Ja. G. Sinai, in *Statistical Mechanics* (Proc. of the I.U.P.A.P. Meeting, Copenhagen, 1966), edited by T. A. Bak (W. A. Benjamin, Inc., New York, 1967).
9. E. C. Percival, *J. Math. Phys.* **2**, 235 (1961).
10. A. Hobson, *J. Chem. Phys.* **46**, 1365 (1957).
11. A. Messiah, *Quantum Mechanics* (John Wiley and Sons, Inc., New York, 1966), Vol. I.
12. J. M. Blatt, *Progr. Theoret. Phys. (Kyoto)* **22**, 745 (1959).
13. J. L. Lebowitz and P. G. Bergmann, *Ann. Phys. (N.Y.)* **1**, 1 (1959).
14. R. C. Tolman, *The Principles of Statistical Mechanics* (Oxford University Press, Oxford, 1938).
15. B. R. A. Nijboer, in *Fundamental Problems in Statistical Mechanics*, edited by E. G. D. Cohen (North-Holland Publishing Company, Amsterdam, 1962).
16. E. T. Jaynes, in *Statistical Physics* (1962 Brandeis Lectures, Vol. 3), edited by K. W. Ford (W. A. Benjamin, Inc., New York, 1963); *Am. J. Phys.* **33**, 391 (1965).
17. H. Mori, *J. Phys. Soc. Japan* **11**, 1029 (1956).
18. K. Kawasaki, *Prog. Th. Phys.* **23**, 754 (1960).
19. D. Scalapino, *Irreversible Statistical Mechanics and the Principle of Maximum Entropy* (PhD. Dissertation, Stanford University, 1961).
20. D. N. Zubarev, *Doklady* **6**, 776 (1962).
21. H. Schwegler, *Z. Naturforsch.* **20a**, 1543 (1965).
22. B. Robertson, *Phys. Rev.* **144**, 151 (1966); *Phys. Rev.* **160**, 175 (1967).
23. A. Hobson and D. N. Loomis, *Phys. Rev.* **173**, 285 (1968).
24. M. Born, *Kgl. Danske Videnskab. Selskab. Mat. Fys. Medd.* **30**, paper 2 (1955).
25. Beauregard, O. Costa de, in *Logic, Methodology and Philosophy of Science*, edited by Y. Bar-Hillel (North-Holland Publishing Co., Amsterdam, 1965).

APPENDIX A

Proof of the Uniqueness Theorem

We will show that the only function satisfying postulates (2.3.4), (2.3.5), (2.3.6), (2.3.7), and (2.3.10) is $k \Sigma p_i \ln (p_i/p_i^0)$. Shannon[1], Feinstein[2], Khinchin[3] and others have proved a similar (but less complicated) uniqueness theorem for Shannon's information measure (2.3.2). Portions of the present proof will follow Feinstein[2] and Khinchin[3].

We first show that

$$I(q_1, \ldots, q_{n-1}, q_n; \ldots) = I(q_1, \ldots, q_{n-2}, q_{n-1} + q_n; \ldots) + (q_{n-1} + q_n) I\left(\frac{q_{n-1}}{q_{n-1} + q_n}, \frac{q_n}{q_{n-1} + q_n}; \ldots\right)^* \tag{A.1}$$

Proof: By postulate (2.3.10), with $Q \equiv q_1 + \cdots + q_{n-2}$,

$$I(q_1, \ldots, q_n; \ldots) = I(Q, q_{n-1} + q_n; \ldots) + QI\left(\frac{q_1}{Q}, \ldots, \frac{q_{n-2}}{Q}; \ldots\right) + (q_{n-1} + q_n) I\left(\frac{q_{n-1}}{q_{n-1} + q_n}, \frac{q_n}{q_{n-1} + q_n}; \ldots\right). \tag{A.2}$$

Again using postulate (2.3.10),

$$I(q_1, \ldots, q_{n-2}, q_{n-1} + q_n; \ldots) = I(Q, q_{n-1} + q_n; \ldots) + QI(q_1/Q, \ldots, q_{n-2}/Q; \ldots) + (q_{n-1} + q_n) I(1; 1). \tag{A.3}$$

By postulate (2.3.6), $I(1; 1) = 0$. Combining (2) and (3), we obtain (1).

Next, we show

$$I(p_{11}, \ldots, p_{1m}, \ldots, p_{n1}, \ldots, p_{nm}; \ldots) = I(q_1, \ldots, q_n; \ldots) + \sum_{i=1}^{n} q_i I(p_{i1}/q_i, \ldots, p_{im}/q_i; \ldots). \tag{A.4}$$

* Concerning the notation: Three dots appearing to the right of the semicolon will always mean that the variables on the right are identical with those on the left, only with the super-script "0" attached. For instance, $I(p_1, p_2; \ldots)$ means $I(p_1, p_2; p_1^0, p_2^0)$.

where $q_i \equiv p_{i1} + \cdots + p_{im}$. *Proof:* By postulate (2.3.10), equation (4) holds for $n = 2$. We proceed by induction. Assuming (4) holds for n,

$$I(p_{11}, \ldots, p_{1m}, \ldots, p_{n+1,1}, \ldots, p_{n+1,m}; \ldots) = I(q_1, \ldots, q_{n-1}, q_n + q_{n+1}; \ldots) + \sum_{i=1}^{n-1} q_i I\left(\frac{p_{i1}}{q_i}, \ldots, \frac{p_{im}}{q_i}; \ldots\right) + (q_n + q_{n+1}) \times \times I\left(\frac{p_{n1}}{q_n + q_{n+1}}, \ldots, \frac{p_{nm}}{q_n + q_{n+1}}, \frac{p_{n+1,1}}{q_n + q_{n+1}}, \ldots, \frac{p_{n+1,m}}{q_n + q_{n+1}}; \ldots\right). \tag{A.5}$$

By (1), the first term on the right hand side of (5) may be written

$$I(q_1, \ldots, q_n, q_{n+1}; \ldots) - (q_n + q_{n+1}) I\left(\frac{q_n}{q_n + q_{n+1}}, \frac{q_{n+1}}{q_n + q_{n+1}}; \ldots\right).$$

By postulate (2.3.10), the last term on the right hand side of (5) may be written

$$(q_n + q_{n+1}) I\left(\frac{q_n}{q_n + q_{n+1}}, \frac{q_{n+1}}{q_n + q_{n+1}}; \ldots\right) + q_n I\left(\frac{p_{n1}}{q_n}, \ldots, \frac{p_{nm}}{q_n}; \ldots\right) + q_{n+1} I\left(\frac{p_{n+1,1}}{q_{n+1}}, \ldots, \frac{p_{n+1,m}}{q_{n+1}}; \ldots\right).$$

Substituting these expressions into the appropriate places in (5), we obtain (4) for $n+1$. Thus (4) holds for every n.

Define

$$f(r, r_0) \equiv I(1/r, \ldots, 1/r, 0, \ldots, 0; 1/r_0, \ldots, 1/r_0).^* \tag{A.6}$$

By (4), with $m = r_0$ and $n = s_0$,

$$I(p_{11} = 1/rs, \ldots, p_{1r} = 1/rs, p_{1,r+1} = 0, \ldots, p_{1r_0} = 0, \ldots, p_{s1} = 1/rs, \ldots, p_{sr} = 1/rs, p_{s,r+1} = 0, \ldots, p_{sr_0} = 0, p_{s+1,1} = 0, \ldots, p_{s_0r_0} = 0; 1/r_0s_0, \ldots, 1/r_0s_0) = I(1/s, \ldots, 1/s, 0, \ldots, 0; 1/s_0, \ldots, 1/s_0) + s(1/s) I(1/r, \ldots, 1/r, 0, \ldots, 0; 1/r_0, \ldots, 1/r_0).$$

* The symbols r, r_0, s, s_0, will always mean positive integers, with $r_0 \geq r$ and $s_0 \geq s$.

Using (6) and postulate (2.3.5), this becomes

$$f(rs, r_0 s_0) = f(s, s_0) + f(r, r_0). \tag{A.7}$$

By (7) and postulate (2.3.6),

$$f(rs, rs_0) = f(s, s_0) + f(r, r) = f(s, s_0). \tag{A.8}$$

Letting $r_0/r = r_0'/r_0 \equiv u$, (8) implies

$$f(r, r_0) = f(r, ur) = f(r, r_0' r/r') = f(r'r, r_0' r) = f(r', r_0'). \tag{A.9}$$

Equation (9) says that $f(r, r_0)$ depends only on the ratio r_0/r. Thus

$$g(r_0/r) \equiv f(r, r_0) \tag{A.10}$$

defines a single-valued function $g(x)$, defined for every rational $x \geqslant 1$. By (10) and (7), defining $x = r_0/r$ and $x' = r_0'/r'$,

$$g(xx') = g(r_0 r_0'/rr') = f(rr', r_0 r_0') = f(r, r_0) + f(r', r_0') = g(r_0/r) + g(r_0'/r').$$

Thus

$$g(xx') = g(x) + g(x') \tag{A.11}$$

for every rational $x \geqslant 1$ and $x' \geqslant 1$. By postulate (2.3.7), $g(x)$ is a monotonic increasing function.

We now show that

$$f(r, r_0) = g(r_0/r) = k \ln (r_0/r) \tag{A.12}$$

where k is a positive constant. *Proof:* By (11),

$$g(r_0^n) = n g(r_0). \tag{A.13}$$

Let integers n, r_0, s_0 be given arbitrarily, and let the integer m be determined by

$$m \leqslant n \ln s_0/\ln r_0 < m + 1 \tag{A.14}$$

from which

$$r_0^m \leqslant s_0^n < r_0^{m+1} \tag{A.15}$$

$$m/n \leqslant \ln s_0/\ln r_0 < (m + 1)/n. \tag{A.16}$$

Since g is monotonic increasing, (13) and (15) imply

$$m g(r_0) \leqslant n g(s_0) \leqslant (m + 1) g(r_0),$$

so that

$$m/n \leqslant g(s_0)/g(r_0) \leqslant (m + 1)/n. \tag{A.17}$$

By (16) and (17),

$$\left| \frac{g(s_0)}{g(r_0)} - \frac{\ln s_0}{\ln r_0} \right| \leqslant \frac{1}{n}. \tag{A.18}$$

Since the left hand side of (18) is independent of n, and since n can be chosen arbitrarily large, (18) implies that

$$g(s_0)/g(r_0) = \ln s_0/\ln r_0$$

or

$$g(r_0) = k \ln r_0. \tag{A.19}$$

Since g is monotonic increasing, the constant k must be positive. This proves (12) for $r = 1$ and any r_0. By using precisely the same reasoning as was used in obtaining (19), it can be shown that

$$g\left(\frac{r+1}{r}\right) = k \ln \left(\frac{r+1}{r}\right). \tag{A.20}$$

(Simply replace r_0 by $(r + 1)/r$ and s_0 by $(s + 1)/s$; equations (13) through (19) all remain valid under this replacement). We now prove (12) by induction on r. Assume (12) is valid for (r_0, r). Then, by (11) and (20),

$$g\left(\frac{r_0}{r}\right) = g\left(\frac{r_0}{r+1}\,\frac{r+1}{r}\right) = g\left(\frac{r_0}{r+1}\right) + g\left(\frac{r+1}{r}\right),$$

$$g\left(\frac{r_0}{r+1}\right) = g\left(\frac{r_0}{r}\right) - g\left(\frac{r+1}{r}\right) = k \ln\left(\frac{r_0}{r}\right) - k \ln\left(\frac{r+1}{r}\right) = k \ln\left(\frac{r_0}{r+1}\right).$$

Hence, if (12) holds for (r_0, r) then it holds for $(r_0, r+1)$. But by (19), equation (12) holds for $(r_0, 1)$. Hence (12) holds for all (r_0, r).

Let p and p^0 be rational numbers less than or equal to 1: $p = r/s \leqslant 1$, $p^0 = r_0/s_0 \leqslant 1$. If the three conditions $r_0 \geqslant r$, $s_0 \geqslant s$, and $s_0 - r_0 \geqslant s - r$ do not hold for the choice (r, s, r_0, s_0), then replace (r_0, s_0) by (nr_0, ns_0), with n sufficiently large that the three conditions hold for (r, s, nr_0, ns_0). This replacement does not affect p or p^0. Hence it may be assumed, without loss of generality, that $r_0 \geqslant r$, $s_0 \geqslant s$, $s_0 - r_0 \geqslant s - r$. By postulate (2.3.5) and equation (4), with appropriate grouping of the variables,

$$\begin{aligned} &I(1/s, \ldots, 1/s, 0, \ldots, 0; 1/s_0, \ldots, 1/s_0) \\ &\quad = I[r/s, (s-r)/s; r_0/s_0, (s_0 - r_0)/s_0] \\ &\qquad + (r/s) I(1/r, \ldots, 1/r, 0, \ldots, 0; 1/r_0, \ldots, 1/r_0) \\ &\qquad + [(s-r)/s] I[1/(s-r), \ldots, 1/(s-r), 0, \ldots, 0; 1/(s_0 - r_0), \ldots, \\ &\qquad\quad 1/(s_0 - r_0)]. \end{aligned}$$

By (6) and (12), this becomes

$$I(p, 1-p; p^0, 1-p^0) = f(s, s_0) - pf(r, r_0) - (1-p)f(s-r, s_0-r_0)$$
$$= kp \ln (p/p^0) + k(1-p) \ln [(1-p)/(1-p^0)].$$

By continuity (postulate (2.3.4)), this result extends to all irrational p. This proves the theorem for $n=2$. Using (1), the theorem follows by induction on n.

REFERENCES

1. See reference 12 of Chapter 2.
2. A. Feinstein, *Foundations of Information Theory* (McGraw-Hill Book Company, Inc., New York, 1958).
3. A. I. Khinchin, *Mathematical Foundations of Information Theory* (Dover Publications, Inc., New York, 1957).

APPENDIX B

Proofs of (3.6.7.) and (3.6.8.)

SINCE THE DATA of (3.6.7) is symmetric, $\varrho(x, 0)$ must be symmetric. Assume that $\varrho^{(1)}(\mathbf{q}, \mathbf{p})$ satisfies the data, and define

$$f(x) = \prod_{i=1}^{N} \varrho^{(1)}(\mathbf{q}_i, \mathbf{p}_i). \tag{B.1}$$

Let $g(x)$ be any other symmetric distribution having $\varrho^{(1)}$ as its reduced distribution:

$$\int g(x)\, d\mathbf{q}_2, \ldots, d\mathbf{q}_N d\mathbf{p}_2, \cdots, d\mathbf{p}_N = \varrho^{(1)}(\mathbf{q}, \mathbf{p}).$$

Then

$$\int g(x) \ln f(x)\, dx = \sum_i \int g(x) \ln \varrho^{(1)}(\mathbf{q}_i, \mathbf{p}_i)\, dx$$
$$= \sum_i \int \varrho^{(1)}(\mathbf{q}_i, \mathbf{p}_i) \ln \varrho^{(1)}(\mathbf{q}_i, \mathbf{p}_i)\, d\mathbf{q}_i\, d\mathbf{p}_i$$
$$= \int f(x) \ln f(x)\, dx.$$

Hence, using (3.6.4) with $h^{3N} N_1! \ldots N_r! \equiv L$,

$$U[f] - U[g] \equiv \int g(x) \ln [Lg(x)]\, dx - \int f(x) \ln [Lf(x)]\, dx$$
$$= \int g(x) \ln \left[\frac{g(x)}{f(x)}\right] dx \equiv I[g; f]$$

(see (2.3.14)). Thus, by (2.3.15), $U[f] \geqslant U[g]$ with equality only if $f(x)$ and $g(x)$ are identical. Thus any distribution $g(x)$ satisfying the data but *not* having the form (1) cannot extremize U. Thus the extremizing distribution must have the form (1).

The result (3.6.8) will be proven only for the case $k = 1$, and for a system having only one degree of freedom; the general result then follows fairly easily. The given information concerns only the phase function $\phi(q, p)$, and hence can be expressed entirely in terms of the probability density ϱ for $\phi(q, p)$. Assuming $\varrho_\phi(\phi)$ satisfies the data, define the distribution

$$f(q, p) = c_1 \varrho_\phi [\phi(q, p)], \tag{B.2}$$

where c_1 is a normalizing constant. We will show that any other distribution $g(q, p)$, having ϱ_ϕ as its reduced distribution for ϕ alone but *not* having the form (2), must have a smaller uncertainty than the distribution defined by (2). It will then follow that, of all distributions satisfying the given information concerning $\phi(q, p)$, the one having the greatest uncertainty has the form (2) and hence is functionally dependent on $\phi(q, p)$.

Let $g(q, p)$ be any other distribution having the same reduced density for ϕ:

$$\varrho_\phi(\phi) = \lim_{\Delta\phi \to 0} \frac{1}{\Delta\phi} \int_{\phi < \phi(q,p) < \phi + \Delta\phi} g(q, p)\, dq\, dp. \tag{B.3}$$

Define a change of variables $(q, p) \to (q', p')$, where

$$q' = \phi(q, p), \quad p' = \theta(q, p), \tag{B.4}$$

where the function $\theta(q, p)$ is chosen in such a way that the Jacobian $J(q, p)$ of the transformation is identically 1. This is always possible; for any given $\phi(q, p)$ the condition $J \equiv 1$ is just a partial differential equation for $\theta(q, p)$. Note that the transformation needn't be canonical, i.e. (q', p') needn't be canonical variables. Under this transformation, the density (2) transforms into

$$f'(q', p') = c_1 \varrho_\phi(q'). \tag{B.5}$$

By (3), using $J \equiv 1$, the transformed density $g'(q', p')$ satisfies

$$\int g'(q', p')\, dp' = \varrho_\phi(q'). \tag{B.6}$$

Using (5), (6) and $J \equiv 1$, we find that

$$\int g(q, p) \ln f(q, p)\, dq\, dp = \int f(q, p) \ln f(q, p)\, dq\, dp.$$

Hence, as in the proof of (3.6.7), $U[f] \geq U[g]$, and the extremizing function must have the form (2). In generalizing this result to the case of arbitrary k and more than one degree of freedom, it should be noted that the required condition $J \equiv 1$ can be fulfilled only if the $\phi_i(x)$ are functionally independent, since otherwise $J \equiv 0$; that is, the functions $\phi_i(x)$ can be used as new variables only if they are functionally independent.

APPENDIX C

Proof that Discrete Classical Mixtures are Almost Periodic

THE PROOF will follow Ref. 10 of Chapter 5. The initial distribution for the discrete mixture "x_k with probability p_k" is

$$\varrho(x, t = 0) = \Sigma\, p_k\, \delta(x - x_k). \tag{C.1}$$

By (3.3.3), the distribution at time t is

$$\varrho(x, t) = \sum_k p_k\, \delta[x - X(t|x_k)] \tag{C.2}$$

where $X(t|x_k)$ is the phase point at t corresponding to the initial point x_k. Expectation values are

$$\langle g(x) \rangle_t = \int g(x)\, \varrho(x, t)\, dx = \sum_k g[X(t|x_k)]\, p_k. \tag{C.3}$$

We will show that, if $g(x)$ is continuous and if $g[X(t|x_k)]$ is bounded for all t and all k (this will certainly hold if $g(x)$ is a bounded function), then the expectation value (3) is an almost periodic function of time.

By the Poincaré recurrence theorem (5.1.3), the phase motion $X(t|x_k)$ is almost periodic, i.e. $X(t|x_k)$ keeps returning arbitrarily near to x_k. Since $g(x)$ is a continuous function, it follows that $g[X(t|x_k)]$ keeps returning arbitrarily near to $g(x_k)$, so that $g[X(t|x_k)]$ is almost periodic. By (3), $\langle g(x) \rangle_t$ is then a sum of almost periodic functions.

A basic theorem of almost periodic functions says that any uniformly convergent sum of almost periodic functions is itself almost periodic. Thus, our proof will be complete if we can show that the sum (3) is uniformly convergent on $(-\infty < t < \infty)$. By assumption, $|g[X(t|x_k)]|$ is bounded by some number M for all (t, k). Thus $|p_k g[X(t|x_k)]|$ is bounded for all t by $M_k = p_k M$. But $\Sigma M_k = \Sigma\, p_k M = M$ is convergent. Thus, by Weierstrass's M test, the sum (3) is uniformly convergent on $(-\infty < t < \infty)$.

Thus, initial distributions which are discrete lead to predictions which are almost periodic.

APPENDIX D

Proof and Discussion of (5.3.2.) (Almost-Periodicity of Quantum Statistical Mechanics)

WE WISH TO PROVE that $\hat{\varrho}(t)$ keeps returning arbitrarily near to $\hat{\varrho}(0)$, where the "distance" $|\hat{\varrho}(t) - \hat{\varrho}(0)|$ is defined by[1]

$$|\hat{A}|^2 \equiv \mathrm{Tr}\,(\hat{A}^\dagger \hat{A}). \tag{D.1}$$

We will follow Percival's proof (Ref. 9 of Chapter 5). Define the operators

$$\hat{\varrho}_{jk}(t) \equiv |H_j\rangle \langle H_j|\, \hat{\varrho}(t)\, |H_k\rangle \langle H_k| \tag{D.2}$$

where the $|H_j\rangle$ are energy eigenvectors. Using (4.3.6),

$$\hat{\varrho}(t) = \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} \hat{\varrho}_{jk}(t)$$

$$= \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} \hat{\varrho}_{jk}(0) \exp(i\omega_{jk} t) \tag{D.3}$$

where $\omega_{jk} \equiv (H_k - H_j)/\hbar$. Thus $\hat{\varrho}(t)$ is a "generalized Fourier series" (i.e. a discrete sum of sine and cosine terms, where the frequencies need not be integral multiples of a fundamental frequency) with operator coefficients.

Any generalized Fourier series which is uniformly convergent, and which has *numerical* coefficients, is almost periodic (see Ref. 1 of Chapter 5). Bochner[2] has generalized this basic result to include series whose coefficients are elements of a metric space. Bochner's generalization applies to (3), where the coefficients $\hat{\varrho}_{jk}(0)$ are elements in the metric space in which distance is defined by (1).[1]

The proof will be complete when we have shown that (3) converges uniformly on $(-\infty < t < \infty)$. Consider the finite sum

$$\hat{\varrho}^{(n)}(t) = \sum_{j=1}^{n} \sum_{k=1}^{n} \hat{\varrho}_{jk}(t)$$

as an approximation to (3). We know that $\hat{\varrho}^{(n)}(t)$ converges to $\hat{\varrho}(t)$ as $n \to \infty$. The question is: does $\hat{\varrho}^{(n)}(t)$ converge *uniformly* to $\hat{\varrho}(t)$ on $(-\infty < t < \infty)$? The square of the "error" is

$$|\hat{\varrho}(t) - \hat{\varrho}^{(n)}(t)|^2 = \left| \sum_{j=n+1}^{\infty} \sum_{k=n+1}^{\infty} \hat{\varrho}_{jk}(t) + \sum_{j=1}^{n} \sum_{k=n+1}^{\infty} \hat{\varrho}_{jk}(t) + \sum_{j=n+1}^{\infty} \sum_{k=1}^{n} \hat{\varrho}_{jk}(t) \right|^2$$

$$= \sum_{j=n+1}^{\infty} \sum_{k=n+1}^{\infty} |\hat{\varrho}_{jk}(t)|^2 + \sum_{j=1}^{n} \sum_{k=n+1}^{\infty} |\hat{\varrho}_{jk}(t)|^2$$

$$+ \sum_{j=n+1}^{\infty} \sum_{k=1}^{n} |\hat{\varrho}_{jk}(t)|^2$$

$$= \sum_{j=n+1}^{\infty} \sum_{k=n+1}^{\infty} |\hat{\varrho}_{jk}(0)|^2 + \sum_{j=1}^{n} \sum_{k=n+1}^{\infty} |\hat{\varrho}_{jk}(0)|^2$$

$$+ \sum_{j=n+1}^{\infty} \sum_{k=1}^{n} |\hat{\varrho}_{jk}(0)|^2.$$

The second equality follows from the orthogonality of the $\hat{\varrho}_{jk}(t)$:

$$\mathrm{Tr}\,(\hat{\varrho}_{j'k'}^\dagger \hat{\varrho}_{jk}) = |\hat{\varrho}_{jk}|^2\, \delta_{jj'} \delta_{kk'}.$$

The third equality follows from $\hat{\varrho}_{jk}(t) = \hat{\varrho}_{jk}(0) \exp(i\omega_{jk} t)$. Thus the error is independent of t, and can be made as small as desired (for *all* t) by taking n sufficiently large. Thus (3) converges uniformly on $(-\infty < t < \infty)$. This proves (5.3.2).

Almost periodicity means that

$$|\hat{\varrho}(\tau) - \hat{\varrho}(0)|^2 = \mathrm{Tr}\,\{[\hat{\varrho}^\dagger(\tau) - \hat{\varrho}^\dagger(0)]\,[\hat{\varrho}(\tau) - \hat{\varrho}(0)]\}$$

$$= \sum_j \sum_k |\langle H_j|\, \hat{\varrho}(\tau)\, |H_k\rangle - \langle H_j|\, \hat{\varrho}(0)\, |H_k\rangle|^2 \tag{D.4}$$

may be made as small as desired simply by choosing the "recurrence time" τ appropriately. Since the summand of (4) is non-negative, all of the matrix elements $\langle H_j|\, \hat{\varrho}(\tau)\, |H_k\rangle$ must *simultaneously* return near to $\langle H_j|\, \hat{\varrho}(0)\, |H_k\rangle$. Intuitively, this implies that all expectation values $\langle \hat{A} \rangle_t$ keep returning near to their initial values. (For a rigorous proof, see Ref. 3 of Chapter 5).

Almost-periodicity depends on the assumptions that (i) $\hat{\varrho}(0)$ has a discrete spectrum (i.e. $\hat{\varrho}(0)$ is a mixture over a *discrete* set of possible states), and (ii) the energy spectrum is discrete. If (i) doesn't hold, then we must use the

eigen-differential formalism (see Ref. 10 of Chapter 5), and the above analysis breaks down. If (ii) doesn't hold, then $\hat{g}(t)$ is no longer the *discrete* sum (3).

REFERENCES

1. U. Fano, *Revs. Mod. Phys.* **29**, 74 (1957).
2. S. Bochner, *Acta Math.* **61**, 149 (1933).

Index