# The advantages of ignorance: Statistical mechanics

With the exception of a few "asides" (about Brownian motion and so on), just about everything that we have discussed in this course so far has rested on the implicit assumption that we have *complete information* about the system(s) to which we are about to apply the relevant physical laws, or at least complete information for the purposes at hand. For example, when applying Newton's laws to the motion of the planets, we assume that we know, or can find out, to any desired degree of accuracy, their current positions and velocities (and can later re-measure these quantities to check our predictions). Of course, there is great deal about the planets we may not know (and certainly Newton did not know) – e.g., their chemical compositions, whether they are inhabited by intelligent beings and so forth; however, this does not matter, because we believe that these unknowns are irrelevant to a prediction of the motion of the planets in the gravitational field of the sun. Similarly, in dealing á la Galileo with the motion of, say, a cannonball here on earth we do not need to know the detailed configuration of each of the atoms composing it (indeed, even gross properties such as chemical composition and overall shape are relevant only to the extent that they influence the effects of air resistance, etc.) Again, in the discussion of special and general relativity, it was only certain very general properties of "events" that played a role: phenomena such as the Doppler effect and the gravitational redshift are completely independent of the detailed nature of the source emitting the radiation! When we come to QM, the question of "complete information" takes on a rather different aspect; according to the usual interpretation of the theory, the most complete information we shall ever in able to get about the individual system is encapsulated in the state vector (wave function) of the ensemble from which that system is drawn. Unlike the case of classical physics, this "complete" information does not permit the prediction with certainty of the outcome of arbitrary subsequent measurements we may make on the system, but within the usual framework it is nevertheless the best we will ever be able to do, even in principle. In discussing the quantum theory we have almost always assumed, explicitly or otherwise, that we do in fact possess a knowledge of the state vector of an appropriate ensemble from which the system can be regarded as drawn. (E.g., in the Schrödinger's cat thought-experiment as we presented it, we assumed that the "triggering" electron is drawn from an ensemble collimated, monochromated, etc., in a given way). Thus there is, as it were, nothing further for us to be ignorant of.

In most real-life applications of physics, however, "complete" information, either classical or QM, is the exception, and a certain amount of ignorance is the norm. Let us, for example, consider a tray of water in the refrigerator and ask how

long it will take to freeze. Without any detailed theory of freezing, we know at least that the water is composed of a very great number of $H_2O$ molecules, and one would certainly imagine that the time in question might depend rather strongly on their initial state. But we certainly do not know this state in detail: if we describe them classically, we do not know the position and velocity of each molecule, which is what is necessary for a complete specification of the initial conditions; and if we describe them (as we in principle should) by QM, we equally certainly do not know the appropriate state vector. In the context of this last remark, it should be emphasized that in applying QM to real-life systems we are faced with, as it were, two kinds of uncertainty. One, which we have already discussed in detail, is associated with the very nature of QM, and in particular with its characteristic "indeterminacy": even if we have the most complete information we are ever allowed to get, in the form of knowledge of the state vector of the relevant ensemble, we still cannot necessarily predict with certainty the outcome of all measurements. The second kind of uncertainty is as it were much more everyday: we simply may not, in practice, know enough about the system in question to be able to assign it reliably to a unique ensemble. In this case, all we can do is to state that with some probability $P_i$ it comes from an ensemble of type i and therefore has a state vector $|\Psi_i\rangle$. (For example, we might have a photon that could have come from either of two beams, one of which has passed through a horizontal polarizer and the other through a vertical polarizer: in this case all we can say is that the photon either has polarization $|\uparrow\rangle$ or polarization $|\rightarrow\rangle$, with probability 1/2 each. This should be carefully distinguished from the quite different statement that the photon has a linear combination of the two polarizations, i.e., is described by the superposition state $\frac{1}{\sqrt{2}}(|\uparrow\rangle + |\rightarrow\rangle) \equiv |\nearrow\rangle$.) This type of "ignorance" is not really in any way peculiar to QM and has no spectacular (interference-type) effects.

In the first half of the nineteenth century, the science of thermodynamics was developed in order to deal phenomenologically with situations, in particular involving macroscopic bodies, where we do not have all the "relevant" information. Later, it was shown how to relate the concepts of thermodynamics to a more first-principles, "microscopic" picture: the resulting branch of physics is known as "statistical mechanics". In this lecture I shall first review classical thermodynamics and then its statistical mechanical underpinnings.

In the course of this exploration we shall run across one very striking difference between thermodynamics and the other branches of physics we have looked at so far. Namely, while classical mechanics, special and general relativity, and even QM (more on this later) are indifferent with respect to interchange of the past and the future, thermodynamics seems to require and impose a definite sense of the "arrow of time" – there really is a difference between the past and the future! We will explore the basis for this difference in the next lecture.

The experience of hot and cold is perhaps one of the most basic known to humankind. With a little more sophistication, we form the concept of "equally hot", and eventually realize, perhaps with a little conscious experimentation, that two bodies put in contact but otherwise effectively isolated from the rest of the world (e.g., the tea and the teapot, given that it is covered with a cozy!) end up being "equally hot". Rather than simply define "equally hot" on the basis of our immediate perceptions (which could be rather dangerous, the tea may actually *feel* hotter than the teapot) we can define two bodies as equally hot, if when they are put in contact, neither gets any hotter or cooler. Being "equally hot" is what logicians call a *transitive relation*: if $A$ is as hot as $B$, and $B$ as hot as $C$, then $A$ is as hot as $C$; this can be regarded as an experimental fact and is sometimes called the "zeroth law of thermodynamics". We can now say that any two bodies which are equally hot possess the same value of some "physical" quantity which is conventionally called the temperature and denoted $T$. Note that at the present stage this is just an abstract definition, and we have no idea how $T$ is related to the other, e.g., mechanical, properties of the body in question. Obviously, how we actually assign a numerical value to $T$ is somewhat arbitrary; we have possibilities ranging from the degree of discomfort felt by our fingers to the height of column of mercury to the volume of a constant-pressure gas thermometer. We will see later that statistical mechanics provides a possible unique, if somewhat abstract, definition. For the moment it will be enough that higher values of $T$ correspond to our intuitive notion of "hotter".

How does a body become "hotter" or "colder"? The most obvious way is by being put in contact with another body that is at a given temperature and transferring "heat" (as we would say in everyday life) to it. At first sight, this looks very similar to the way in which, when we put a positively charged body in contact with a negatively charged one, electric charge flows between them. It is therefore at first sight tempting to regard 'heat" as like electric charge, that is, as a substance of which a body possesses a certain amount, so that the total amount of "heat" is conserved. This view was popular in the eighteenth century, and the substance in question was known as "caloric". However, it had been known qualitatively for a long time, and was verified quantitatively in a famous (cannon boring) experiment by Count Rumford in 1796, that it is possible to raise the temperature of a thermally isolated body by doing mechanical work on it. Thus, if "caloric" exists at all, it is certainly not conserved! In fact, a given amount of mechanical work is always equivalent to the same amount of heat, irrespective of the nature of the substance. A (thought!) experiment to demonstrate this might go as follows: Suppose that when a given quantity of water at say 20°C and of brass at 50°C are put in thermal contact, they eventually come into equilibrium (i.e., become "equally hot") at 30°C. By definition, the amount of "heat" lost

by the brass is equal to that gained by the water. Thus it will require an equal amount of mechanical work (under ideal conditions, of course!) to re-heat the brass to 50°C and to re-heat the water to 30°C after it has been re-cooled to 20°C. Thus, heat and mechanical work are nothing but different forms of the same commodity – energy! The only problem is that, because of the history, they are conventionally measured in different units: work is measured in joules, but heat in calories, one calorie being the heat require to raise the temperature of 1 gram of water by 1 degree, under specified conditions. The equivalence is: 4.2 joules equals one calorie. (In everyday terms, this ratio is enormous, as anyone who has spent a night at an Alpine hut above the snow line and tried to use mechanical work to convert snow into drinking water knows!)

The equivalence of heat and mechanical work as different forms of energy, plus the general principle (which we tend to believe on rather general grounds) that total energy is conserved, constitute what is known as "the first law of thermodynamics": crudely speaking "work + heat = constant". E.g., if we imagine that we carry a thermodynamic system (such as the freon in a household refrigerator) around a cycle of operations, returning it at the end to its original state so that its total energy is unchanged, then we know that the total heat input to the system over the cycle must be equal to the mechanical work output (e.g., in the expansion and compression of the gas); however, each of these quantities can be individually nonzero (and must be if the refrigerator is to work!).

Nothing so far explains why heat tends to flow from a hotter to a cooler body but not vice versa. Evidently, it would be nice if on a cold winter's day we could simply pull in energy, in the form of heat, from the outside air and use it to heat the house! In fact, nothing in the first law of thermodynamics forbids us doing so, and modern "heat pumps" do achieve this end. However, to do so they have to perform mechanical work; if no such device is involved, we know very well that not only can we not do this, but in fact the reverse will happen (the house, without continuous input of heat from the furnace, will eventually cool to the temperature of the outside air). Reflections on these and similar "impossibilities" led Carnot, Clausius and others in the early 19th century to arrive, by a series of very elegant arguments, at the concept of entropy and the related second law of thermodynamics. The idea is that in thermal equilibrium under specified conditions (given temperature, given pressure, etc.) any physical system will possess not only a definite total energy but also a definite value of another quantity, called "entropy" and usually denoted $S$.* Although the "total amount of heat" possessed by the body is a meaningless concept (since heat alone is not conserved) the *heat input* $\Delta Q$ (which is well-defined) can be related to the *change* $\Delta S$ of the entropy: if the input is gradually

*Like the energy (but unlike, e.g., the temperature) entropy is an "extensive" quantity, i.e., half the body has half the entropy (but not, of course, half the temperature!).

done while the body is at a given temperature $T$ and "nothing else" happens to the system (e.g., no cannon-boring, etc.) then the relation is $\Delta S = \Delta Q/T$. The second law of thermodynamics can be expressed in various forms that turn out to be equivalent; in the present context it is convenient to express it by the statement that, for a thermally isolated system or set of systems (i.e., such that there is no external heat input), the entropy can only increase or, at best, remain constant: $\Delta S_{\rm tot} \geq 0$. It is immediately clear how this principle ensures that the transfer of heat will be from the hotter body to the cooler and not vice versa (assuming that no third body is involved, as will be the case if a heat pump is applied). Suppose bodies 1 and 2 are initially at temperatures $T_1$ and $T_2$ respectively, and are put into thermal contact while thermally isolated from the rest of the world (e.g., tea and teapot). Let $\Delta Q$ be the heat lost by body 1 and therefore gained by body 2; for the moment we allow $\Delta Q$ to have either sign. From the above we have $\Delta S_1 = -\Delta Q/T$, $\Delta S_2 = +\Delta Q/T_2$, and hence the total entropy change of the two bodies combined is $\Delta S_{\rm tot} = \Delta Q(1/T_2 - 1/T_1)$. But the second law says that $\Delta S_{\rm tot}$ must be $\geq 0$. Since by construction, $T_1 > T_2$ and hence $1/T_2 > 1/T_1$, it follows that $\Delta Q$ must be positive, i.e., the heat flows from the hotter to the colder body.

In the above we said "the entropy must *increase*" (or at best stay constant). Let's make this more explicit. "The later the time, the larger the entropy must be." Thus, the second law (unlike the first) picks out certain times as "later" than others, i.e., it defines an *arrow of time*. Processes that involve zero change in total entropy are called "reversible", others "irreversible". Thus, as the names imply, a reversible process can "go" in either direction (the statement "remains constant" does not distinguish past from future!) but *an irreversible thermodynamic process can proceed only in one direction*. Most of the physical processes we observe in everyday life are irreversible, which is why we can usually tell when the movie is being run backwards!

Once we have the definition of "entropy" and the first and second laws, we can derive a large number of results that are quite independent of the detailed nature of the bodies under consideration (e.g., we can relate the compressibility, thermal expansion and specific heat). If, further, we know the way in which the energy and entropy depend on other "thermodynamic" variables such as temperature, pressure, etc., we can do quantitative calculations and start building useful things such as internal combustion engines, refrigerators, etc. This is the subject matter of classical thermodynamics.

It is important to note that to do classical thermodynamics all we need to know are a few macroscopic (and relatively easily measurable) properties of the system (e.g., pressure, volume, temperature). The subject is therefore a completely self-contained one, and indeed in the late nineteenth century there were plenty of people (including Mach, whom we have met before) who argued that there was

simply no need for any more microscopic basis for it. However, the majority of physicists felt a certain discontent; while the first law could perhaps be understood at least qualitatively if we identify the part of the energy called "heat" with random motion of individual atoms/ molecules, which is on too small a scale to manifest itself as macroscopic work, it remained a mystery why the second law holds and in particular what the nature of the mysterious "entropy" is. This discontent was (largely) removed with the development of statistical mechanics, to which I now turn. I will frame the discussion primarily in terms of classical statistical mechanics – i.e., a theory that assumes that the dynamics of individual atoms obeys Newton's rather than Schrödinger's equation – and bring in the quantum version only when it actually simplifies matters.

To illustrate the basic ideas of statistical mechanics, let's start with a specific example. Imagine that we have a very large lattice ($N$ points) with one small magnet ("spin") on each site that can point only "up" or "down". (This is actually not a bad model of certain kinds of magnetic material.) A point that is very crucial in our argument is that it is very unlikely that we shall be able to know the direction of each individual magnet – in fact, it is obviously impossible to find this by any macroscopic measurement. On the other hand, it may be quite easy to find out the total magnetization $M$ of the sample (i.e., the number of "up" spins [magnets] minus the number of "down" spins), e.g., from the magnetic field produced by it; it should also not be too difficult to find the total magnetization of "sufficiently large" parts of the system. For definiteness, let's divide the system into two equal halves; we then assume we can measure $M$, and also $M_1$ and $M_2$, the magnetizations of the two halves, separately. Furthermore, let's assume, for simplicity, that the total energy is a unique function of $M$ and that the system is thermally isolated, so that its energy is fixed at a particular value; then $M$ must also have some fixed value, which for simplicity we choose to be zero. The question is: given this information, what behavior do we predict for the "macroscopically observable" quantities $M_1$ and $M_2$ $(= -M_1)$?

It is clear that, since by hypothesis we are not allowed to know the behavior of individual spins, any predictions we can make must be based on probabilistic considerations. How can we apply probability theory to this problem? We must first decide how to assign an a priori probability to each of the $2^N$ possible configurations of the $N$ spins, or actually more precisely to that subset of them that corresponds to $M = 0$ (since we already have this piece of information). At this point, we remark that there is nothing that distinguishes any one configuration physically from any other,[†] and then use the principle of indifference to argue that therefore, it is legitimate to assign exactly the some a priori probability to each.

[†]In a more realistic model, the energy is very likely to be different for different configurations, but we have got rid of this complication by our assumption that $E$ is a unique function of $M$.

Once we have made this crucial assumption, the rest of the calculation is in principle straightforward: it is clear that the probability of obtaining a given value of $M_1$ (and hence the opposite value of $M_2$) is simply the number of different configurations with $M = 0$ that give this value of $M_1$. In fact, it is clear that the problem is exactly the same as the following one: given that in a long series of coin tosses the total number of heads and tails is equal, what is the probability of getting more heads than tails on the first half of the sequence? (heads $\Rightarrow$ spin up, tails $\Rightarrow$ spin down). Without detailed calculation, we can see what the qualitative behavior will be (cf. lecture 18): first, consider the probability of the "extreme" result that $M_1$ is simply $N/2$. For this to be so, all the spins on side 1 must be up, and there is only one configuration corresponding to this arrangement. What about the case $M_1 = N/2 - 1$? Now there are $N/2$ possible configurations, corresponding to the different possible choices for the single "down" spin on side 1. (Similarly, on side 2 there are $N/2$ possible choices for the single "up".) Thus, the total number of configurations is $(N/2)^2$. For $M_1 = N/2 - 2$, there are $\left[\frac{1}{2}\frac{N}{2}\left(\frac{N}{2} - 1\right)\right]^2$ configurations, and so on. It is intuitively plausible (and turns out to be true) that the most probable value of $M_1$ (and hence also of $M_2$) is zero. It is somewhat less obvious (but follows from a standard calculation) that the value of $M_1$ for which the probability falls to 1/2 of the maximum value is of order $N^{1/2}$. Thus, we would conclude that the most probable value of $M_1$ is zero, and that the "fluctuations" of $M_1$ around this value, expressed as a fraction of the maximum possible value $N/2$, are of order $N^{-1/2}$ (compare lecture 18). Thus, in the limit $N \to \infty$ (usually called the "thermodynamic limit"), the value of the macroscopic quantity $M_1$ is predicted to have a perfectly well-defined value (in this case, zero). Under appropriate circumstances, this prediction is experimentally verified.

The method illustrated in the above "toy" problem can be generalized to any problem involving a system thermally isolated from the outside world. In that case, the energy is fixed, so one considers all those configurations that correspond to the given value of the energy, asks how many correspond to a particular value of the other "thermodynamic" variables (e.g., pressure, voltage) one is interested in predicting, and assigns a probability proportional to the number of configurations. Quite generally, one finds that in the thermodynamic limit (size of system $\to \infty$), the probability peaks very sharply about a particular value,[‡] with fluctuations of order $N^{-1/2}$. Of course, the "counting" of the relevant configurations and hence the application of the principle of indifference may not always be entirely trivial, and sometimes in the last resort we have to fall back on quantum mechanics to justify the counting method we believe (essentially on experimental grounds, as long as we stay with classical physics!) to be correct. But the basic principle is just that illustrated above.

[‡]Or, occasionally, two or more different values (the case of so-called "broken symmetry").

Suppose now that we have actually measured the value of $M_1$ and found that it is *not* the "most probable" value zero, but (say) is of order $N/2$. We would be a little surprised, but might account for this result by, for example, the assumption that our colleague had applied a magnetic field to side 1 (and an opposite one to side 2) so as to orient the spins, and then turned it off just before we walked into the room to do the measurement (cf. next lecture). We now suppose that we leave the system to itself, and that the microscopic dynamics are such that the spins can "flip-flop", that is, exchange spin while conserving the total spin of the sample (and hence, according to our model, the total energy: this is a not unreasonable model for the dynamics of many magnetic systems). These "flip-flops" can recur not only within regions 1 and 2, but also across the 1-2 boundary, in which case they will obviously change $M_1$. Under these circumstances, how do we expect the quantity $M_1$ to evolve in time?

Strictly speaking, the only totally rigorous answer to this question would be an explicit calculation, either by hand or (more likely nowadays) by computer. For such a calculation, we would need as input the exact dynamics, e.g., the probability of a flip-flop transition of each pair of opposite spins in the sample, etc. Furthermore, we would need to do the calculations separately for each of the different states that could have given rise to the initially observed nonzero value of $M_1$, which, unless this value is very close to $N/2$, will be extremely numerous. However, most physicists would probably have a very strong prejudice that unless the dynamics, and/or the initial state, is "pathological", the system would eventually approach the value of $M_1$ (in this case zero), which was indicated by the original calculation based on "complete ignorance", and then stay there. In many relevant cases, that does seem to be the behavior that is observed experimentally.

Thus, at least so long as our system is thermally isolated, there seems to be a characteristic of the macroscopic state (defined by the value of $M_1$ and nothing else) that is a maximum when we have no information about the system at all and that, when we start from a state not corresponding to the maximum, tends to approach the state which does, i.e., to increase. This quantity is the number of configurations that correspond to that value of $M_1$. Now the (macroscopic) second law of thermodynamics tells us that the entropy $S$ tends, for a thermally isolated system, to increase and thus to reach its maximum value. It therefore becomes plausible to associate *the thermodynamic concept, entropy* ($S$), with the statistical concept of *the number of "available" configurations* ($W$), and in fact this association lies at the heart of the discipline now known as "statistical mechanics". However, it is clear that the entropy cannot be simply proportional to $W$, if for no other reason than that when the system is doubled in size without changing, e.g., the mean magnetization, the entropy is doubled (see above) while the number of available configurations is (approximately) squared. In fact, the

only law of association that will get this feature right is to make $S$ proportional to the logarithm of $W$:

$$S = k_{\rm B} \ln W$$

(i.e., in our toy model $S(M_1) = k_{\rm B} \ln W(M_1)$). The constant $k_{\rm B}$ is necessary because $S$ has the units of energy/ temperature ($\Delta S = \Delta Q/T$!); it turns out to have the numerical value of $1.38 \times 10^{-23}\,{\rm J/K}$, and is known as "Boltzmann's constant".

Once we have the above "statistical-mechanical" definition of "entropy", it is relatively straightforward to derive the general formal results quoted in textbooks of statistical mechanics; if we know enough about the dynamics of a specific system, etc., we may also actually be able to calculate the entropy as a function of energy, magnetization, etc., for this system. One important generalization should be noted: we often want, in real life, to consider systems that are not thermally isolated but are in contact with some thermal "bath" (i.e., a very large environment, such as Earth's atmosphere, such that the effects of the system on it are negligible). In general, what tends to be specified is not the energy of the bath (which it can continually exchange with the system, under these "thermal contact" conditions), but rather its "temperature", and so we need to know how "temperature" is to be defined in a statistical-mechanical approach. It turns out that if we define a quantity $T$ by

$$T = \frac{1}{\Delta S/\Delta E}$$

the quantity so defined has all the properties expected of the thermodynamic "temperature", and all the generic results obtained from statistical mechanics can be put into exact one-one correspondence with those of thermodynamics. Whether this means that thermodynamics is "equivalent" to statistical mechanics is a question beloved of philosophers of science (cf., e.g., Sklar pp. 106-8).

One point about the above discussion of entropy cannot be overemphasized: it makes no sense to talk about the entropy of a system if its exact microscopic state is known. "Entropy" is an essentially macroscopic concept, and is defined for a system when we have, at best, knowledge of a few macroscopic quantities; e.g., in the above toy model, the entropy $S(M_1)$ is a property of the macroscopic state characterized by a particular value of $M_1$ (as we saw, it is essentially the logarithm of the number of micro-states that correspond to this value).

What I have outlined above is essentially what is usually called the "information theoretic" approach to entropy; it clearly fits in well with (and could be regarded as a special case of) the "subjectivist" approach to probability in general. There exists another approach, usually called the "ergodic" approach, which can be regarded as in some sense a special case of the "objectivist" or "frequentist" approach to probability. It is based on the so-called "ergodic hypothesis", and closely related

hypotheses such as mixing, K-system behavior, etc. Crudely speaking, in the present context, we can define these as stating that a system with a given fixed energy will in time explore all configurations available to it that are consistent with that constraint and moreover spend equal time in each. Needless to say, this relies as above on having a correct count of the different configurations.

If true, the ergodic hypothesis would do two things for us. First, since the probability of our observing a given microstate is presumably proportional, under normal conditions, to the time the system spends in it, it would justify the use of the principle of indifference used in the information-theoretic approach. Second, it would justify our "prejudice" that the system, when started from a nonequilibrium macrostates, would eventually attain the equilibrium one, since in effect it starts from one of a very small set of configurations (that corresponding to the original nonzero value of $M_1$) and ends up spending the overwhelming majority of its time in a much larger set (that corresponding to $M_1 = 0$). The very small fraction of its time it spends in the long run in the macrostate corresponding to the original (improbable) value of M, would not in practice be observable experimentally.

Unfortunately, not only has the ergodic hypothesis (and related ones) been proved only for a very small set of systems, it has actually been *disproved* for a considerably larger class. The most common reason for its failure, in real life, is the existence of some conservation law or laws different from that of conservation of energy. To take a trivial case, it is clear that if in our toy model we forbid "cross-boundary" flip-flops, thereby conserving $M_1$ and $M_2$ separately, no relaxation to equilibrium can occur. In response to this consideration one might perhaps object that no conservation law is ever completely exact, and that therefore it is only a matter of how long we are prepared to wait. However, a more surprising result that has come out of the (very sophisticated) work of the last few decades in nonlinear mechanics is that in many systems, despite the absence of any particular conservation law for macroscopic quantities (other than the energy, of course), there are often nevertheless classes of configurations that never "communicate" with one another; i.e., if we start in one class, the dynamics will never take us into the other class, in contradiction to the ergodic hypothesis. This result, stated in more quantitative terms, is the famous KAM theorem; for a further discussion, see Sklar, pp. 125-7.

By and large most practicing physicists (as distinct from many mathematicians and philosophers of science!) do not worry over much about the KAM theorem and related disproofs of the ergodic hypothesis, probably because they are conscious that the idealized dynamics used in the proofs are themselves unlikely to correspond exactly to any real-life system. Most physicists in fact probably rather tend to assume that in the absence of macroscopic conservation laws (which there are standard techniques for taking into account) the ergodic hypothesis is true FAPP

– "for all practical purposes". It is unclear whether this spirit of nonchalance will some day let us down....