THE PHYSICS OF INFORMATION

INFORMATION THEORY IN THE LIGHT OF THERMODYNAMICS, STATISTICAL MECHANICS AND NONLINEAR DYNAMICS

DRAFT: JULY 20, 2005

F. ALEXANDER BAIS AND J. DOYNE FARMER

Contents

Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501.

1. The Physics of Information

> Why cannot we write the entire 24 volumes of the Encyclopedia Brittanica on the head of a pin?
>
> R.P. Feynman

Information is carried, stored, retrieved and processed, by machines, whether they be electronic computers or living organisms. The basis for any means of handling information is physical and it is therefore not surprising that physics and information have a rich interface. All information is ultimately carried by a physical substrate, be it paper, silicon chips or holograms and therefore we know that our strings of zeros and ones will have to obey the fundamental laws of physics. In our quest for more and more volume and speed in storing and processing information we are naturally led to the smallest scales we can physically manipulate, and as ultimately all of matter is composed of atoms, the laws of quantum mechanics come into play. We refer to Feynman's visionary 1959 lecture "Plenty of room at the bottom" [4] where he already talks about storing and manipulating information on the atomic level. However, the interface between physics and information is not limited to the hardware implementation of memory and information processing. It also involves a common history in the theoretical domain where it comes to the perception, analysis and understanding of some of the very basic concepts in information theory. In this brief review we focus on the various subfields of physics in which the notion of information or entropy is of paramount importance and we'll highlight some particular examples. Our strategy is to start from the (theoretical) physics side and link the relevant concepts to their information scientific counterparts.

The logical structure of the chapter is as follows. We begin by describing the origin of the concept of entropy in thermodynamics. We discuss how the microscopic theory of atoms led to statistical mechanics, which makes it possible to derive and extend thermodynamics. This led to the definition of entropy in terms of probabilities and provided the inspiration for modern information theory. A close examination of the foundations of statistical mechanics and the need to reconcile the probabilistic and deterministic views of the world leads us to a discussion of chaotic dynamics, where information plays a crucial role in quantifying predictability. We then discuss a variety of fundamental issues that emerge in defining information and how one must exercise care in discussing concepts such as order, disorder, and incomplete knowledge. We also discuss an alternative form of entropy and its possible relevance for nonequilibrium thermodynamics. Toward the end of the chapter we make some excursions into cosmology and engineering. One is the "ultimate information paradox" posed by the physics of Black Holes, the other is an example of how the notion of information is used in an axiomatic approach to design engineering.

In this review we have limited ourselves and not all relevant topics that touch on both physics and information have been covered, notably the subject of quantum information is not treated.

2. Thermodynamics

> The truth of the second law is , therefore, a statistical and not a mathematical truth, for it depends on the fact that the bodies we deal with consist of millions

of molecules and that we never can get a hold of single molecules

J.C. Maxwell

Thermodynamics is the study of macroscopic physical systems[1]. These systems contain a large number of degrees of freedom, typically of the order of Avogadro's number, i.e. $N_A \approx 10^{23}$. The three laws of thermodynamics describe processes in which systems exchange energy with each other or with their environment. For example, the system may do work, or exchange heat or mass through a diffusive process. A key idea is that of *equilibrium*, which in thermodynamics is the assumption that the exchange of energy or mass between two systems is the same in both directions; this is typically only achieved when two systems are left alone for a long period of time. A process is *quasistatic* if it always remains close to equilibrium, which also implies that it is *reversible*, i.e that the process can be undone and the system can return to its original state. It may also be that a system goes from one equilibrium state to another via a nonequilibrium process (think for example of the free expansion of a gas, or the mixing of two fluids), in which case it is not reversible. No real system is fully reversible, but it is nonetheless a very useful concept.

The remarkable property of systems in equilibrium is that the macro states can be characterized by only very few variables such as volume V, pressure P, temperature T, entropy S, chemical potential μ and particle number N. Moreover, these state variables are in general not independent, but rather are linked by an *equation of state*. A familiar example the ideal gas law $PV = N_A kT$, where k is the Boltzmann constant relating temperature to energy ($k = 1.4 \times 10^{-23}$ $joule/sec.$). In general the state variables come in pairs, one of which is *intensive* (like $P, T, ...$) while the other, conjugate variable is *extensive*[2] ($V, S, ...$). The formulation of thermodynamics owes its diversity to the fact that the variables are not independent so that according to the physical setting a suitable choice for the independent variables should be made. In this lightning review we will only highlight the essential features which are most relevant in connection with information theory.

2.1. **The laws.** The first law of thermodynamics reads[3]

$$dU = đQ - đW \tag{2.1}$$

and amounts to the statement that heat is a form of energy and that energy is conserved. More precisely, the change in internal energy dU equals the amount of heat $đQ$ absorbed by the system minus the work done by the system, $đW$.

The second law introduces the concept the entropy S, which is defined as the ratio of heat flow to temperature. It states that the entropy for a closed system (a system with constant energy, volume and number of particles) can never decrease. In mathematical terms:

$$dS = \frac{đQ}{T} \qquad \frac{dS}{dt} \geq 0 \tag{2.2}$$

[1]Many details of this brief expose of selected items from thermodynamics and statistical mechanics can be found in standard textbooks on these subjects [19, 13, 8, 16].

[2]Extensive in this context means proportional to system size, whereas intensive means independent of system size.

[3]The bars through the differentials indicate that the quantities following them are not state variables: the d-bars therefore refer to small quantities rather then proper differentials

By using a gas as the canonical example, we can rewrite the first law in proper differentials as

$$dU = TdS - PdV. \tag{2.3}$$

It follows from the relation between entropy, heat and temperature that entropy differences can in principle be measured by determining the temperature and the change in heat using a thermometer and a calorimeter. This confirms the statement made at the start that thermodynamics involves measurements of macroscopic systems only.

There are two, different formulations that make clear what the second law actually means. The Kelvin formulation states that it is impossible to have a machine whose sole effect is to convert all absorbed heat into work, while the formulation due to Clausius says that it is impossible to have a machine that only extracts heat from a reservoir at low temperature and delivers that same amount of heat to a reservoir at higher temperature. Rephrasing these formulations, Kelvin says that ideal engines cannot exist and Clausius says that ideal refrigerators can't exist. A more modern formulation of the second law, which in the setting of statistical

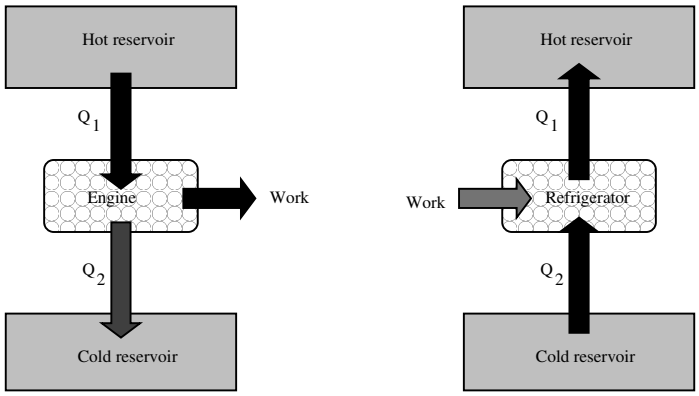

FIGURE 1. The relation between heat and work illustrating the two formulations of the second law of Thermodynamics. On the left we have the Kelvin formulation. The ideal engine corresponds to the diagram with the black arrows only. The second law tells us that the third grey arrow is necessarily there. The right picture with only the black arrows corresponds to the ideal refrigerator, and again the third grey arrow is again a consequence of the second law.

mechanics is equivalent to the statements of Kelvin and Clausius, it is the so-called "Landauer principle", which says that there is no machine whose sole effect is the erasure of information. There is a price to forgetting: the principle states that the erasure of information (which is irreversible) is inevitably accompanied by the generation of heat, one has to generate at least $kT \ln 2$ to get rid of one bit of information [14, 15]. The second law sets fundamental limits on the possible efficiency of real machines like steam engines, refrigerators and information processing devices. As everybody knows, real engines give off heat and real refrigerators and real computers need power to do their job. The second law tells us to what extent

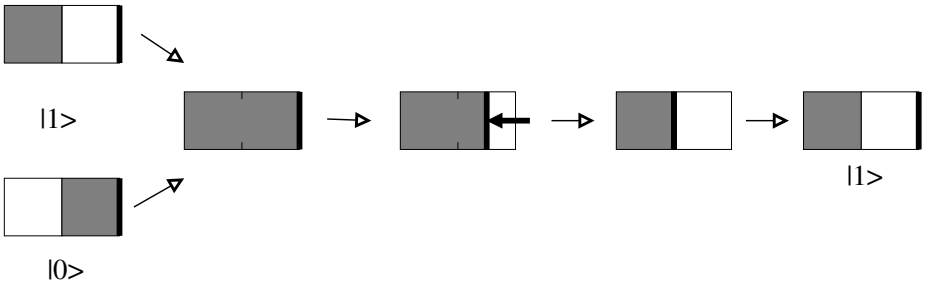

FIGURE 2. An illustration of the Landauer principle, using a thermodynamical system. We consider a "gas" consisting of a single atom in a symmetric container in contact with a heat bath. The atom can be either on the left or on the right, corresponding exactly to one bit of information. Erasing the information amounts to resetting the device to the $|1>$ state independent of the initial state. This can be done by first opening a diaphragm in the middle, then reversibly moving the piston from the right in, and finally closing the diaphragm and moving the piston back. In the first step the gas expands to the double volume. The particle doesn't do any work, the energy is conserved and therefore no heat will be absorbed from the reservoir. In the (quasistatic and isothermal) processes we bring the system back to a state which has the same entropy as the initial state. During the compression the entropy has decreased by $k \log 2$. The conclusion is that the system must have gotten rid of an amount of heat that equals $kT \log 2$. which is exactly the amount of work done on the system by moving the piston, as the energy of the system has not changed.

absorbed heat can be used to perform work. The increase of entropy as we go from one equilibrium situation to another is related to dissipation and the production of heat, which is intimately linked to the important notion of *irreversibility*. A given action in a closed system is irreversible if it makes it impossible for the system to return to the state it was in before the action took place. Irreversibility is always associated with production of heat, because heat cannot be freely converted to other forms of energy (whereas any other form of energy can always be converted to heat). Irreversibility implies path dependence.

The theory of thermodynamics taken by itself does not connect entropy with information. This only comes about when the results are interpreted in terms of a microscopic theory, in which case temperature can be interpreted as being related to uncertainty and incoherence in the position of particles. This requires a discussion of statistical mechanics, as done in the next section.

There is another fundamental aspect to the second law which is important from an operational as well as philosophical point of view. A profound implication of the second law is that it defines an "arrow of time", i.e., it allows us to distinguish the past from the future. This is in contrast to the fundamental laws of physics which are (except for a few exotic interactions) time reversal invariant. At the microscopic level of fundamental particles, if one watches a movie it is very difficult to tell

whether it is running forwards or backwards, except for a few exotic interactions that are only very rarely seen under normal conditions as we find them on earth. In contrast, if we watch a movie of macroscopic events, we do not have to look hard to find irreversible actions such as the curling of smoke, the spilling of a glass of water, or the mixing of bread dough, which easily allow us to determine whether we are running in forward or reverse. More formally, even if we didn't know which way time were running, we could pick out some systems at random and measure their entropy at times $t_1, t_2, ...$ The direction in which entropy increases is the one that is going forward in time. Note that we didn't define an a priori direction of time in formulating the second law. Thus it establishes a time direction on its own, without any reference to atomic theory or any other laws of physics.

We note that the second law of thermodynamics talks only about the difference between the entropy of different macro states. The absolute scale for entropy is provided by the third law of thermodynamics. This law states that when a system approaches the absolute zero of temperature the entropy will go to zero, i.e.

$$T \to 0 \quad \Rightarrow \quad S \to 0 \tag{2.4}$$

When $T = 0$ the heat is zero, corresponding to no atomic motion, and the energy takes on its lowest possible value. We know that such a lowest energy "ground" state exists due to the quantum mechanical nature of matter, but it is interesting that this was already evident from thermodynamics, without any reference to atoms.

Let us conclude by emphasizing that the laws of thermodynamics have a wide applicability and a rich phenomenology that supports them unequivocally.

2.2. **Free energy.** Physicists are particularly concerned with what is called the (Helmholtz) *free energy*, denoted F. It is a quantity that is relevant when studying systems in thermal contact with a heat bath. Furthermore the free energy plays a central role in establishing the relation between thermodynamics and statistical mechanics as we will discuss in the next section.

The free energy is defined by:

$$F \equiv U - TS \ . \tag{2.5}$$

This implies that in differential form we have

$$dF = dU - TdS - SdT \tag{2.6}$$

which, using (2.3) can be written as

$$dF = -PdV - SdT. \tag{2.7}$$

The natural independent variables to describe the free energy are evidently volume and temperature.

Let us briefly reflect on the meaning of the free energy. Consider a system A in thermal contact with a heat bath A' kept at a constant temperature T_0. Suppose the system A absorbs an amount $đ Q$ from the reservoir. Clearly we may think of the total system consisting of system plus bath as a closed system: $A^0 = A + A'$. For A^0 the second law implies that its entropy can only increase: $dS^0 = dS + dS' \geq 0$. As the temperature of the heat bath A' is constant and its "absorbed" heat is $-đQ$, we may write $dS' = -đ\, Q/T_0$. From the first law applied to system A we obtain that $-đ\, Q = -dU - đ\, W$, so that we can rewrite the inequality for the entropy

change as $-dU + T_0 dS \geq đW$ As the system A is kept at a constant temperature the left hand side is just equal to $-dF$ so that we arrive at the inequality

$$-dF \geq đW. \tag{2.8}$$

The maximum work that can be done by the system in contact with a heat reservoir is $(-dF)$. If we keep the system parameters fixed i.e. $đW = 0$ we obtain that $dF \leq 0$, showing that for a system coupled to a heat bath the free energy tends to decrease, and consequently in an thermal equilibrium situation the free energy will acquire a minimum. This statement is to be compared with the statement that for an isolated system in equilibrium the entropy acquires a maximum.

We can think of the second law as telling us how different kinds of energy are converted into one another: In an isolated system, work can be converted into heat, but heat cannot be converted into work. From a microscopic point of view forms of energy that are "more organized", such as light, can be converted into those that are "less organized", such as the random motion of particles, but the opposite is not possible.

From Equation (2.7) the pressure and entropy of a gas can be written as partial derivatives of the free energy

$$P = \left(\frac{dF}{dV}\right)_T \qquad S = \left(\frac{dF}{dT}\right)_V \tag{2.9}$$

So we see that for a system in thermal equilibrium the entropy is a state variable, meaning that if we reversibly traverse a closed path we will return to the same value (in contrast to other quantities, such as heat, which do not satisfy this property). The variables P and S are dependent variables, this can be moist easily seen in the so called Maxwell relation that is obtained by equating the two second derivatives

$$\frac{\partial^2 F}{\partial T \partial V} = \frac{\partial^2 F}{\partial V \partial T} \tag{2.10}$$

yielding the relation

$$\left(\frac{\partial P}{\partial T}\right)_V = \left(\frac{\partial S}{\partial V}\right)_T . \tag{2.11}$$

3. Statistical mechanics

> In dealing with masses of matter, while we do not perceive the individual molecules, we are compelled to adopt what I have described as the statistical method of calculation, and to abandon the strict dynamical method, in which we follow every motion by the calculus.
>
> J.C. Maxwell

> We are forced to be contented with the more modest aim of deducing some of the more obvious propositions relating to the statistical branch of mechanics. Here there can be no mistake in regard to the agreement with the facts of nature.
>
> J.W. Gibbs

Statistical mechanics is the explanation of the macroscopic behavior of physical systems using the underlying microscopic laws of physics even though the microscopic states (such as the position and velocity of individual particles) are unknown. The key figures in the late 19th century development of statistical mechanics were Maxwell, Boltzmann and Gibbs [17, 2, 6]. One of the outstanding questions was to

derive the laws of thermodynamics, in particular to give a microscopic definition of the notion of entropy. Other objectives were the understanding of transport phenomena and transport coefficients from the underlying dynamics, which cannot be computed from thermodynamics alone. For our purpose, which is highlighting the links with information theory, we will give a brief and somewhat lopsided introduction to some of the relevant concepts. Our main goal is to show the origin of the famous expression for the entropy, $S = -\sum_i p_i \ln p_i$, which was later used by Shannon to define information.

3.1. **Definitions and postulates.**

> Considerable semantic confusion has resulted from failure to distinguish between prediction and interpretation problems, and attempting a single formalism to do both.
>
> T.S. Jaynes

Statistical mechanics considers systems with many degrees of freedom, such as atoms in a gas or spins on a lattice. We can think in terms of the microstates of the system which are, for example, the positions and velocities of all the particles in a box of gas. The space of possible microstates is called the *phase space*; for a monatomic gas with N particles, the phase space is $6N$-dimensional, corresponding to the fact that under Newtonian mechanics there are three positions and three velocities that must be measured for each particle in order to determine its future evolution. A microstate of the whole system thus corresponds to a single point in this phase space.

Statistical mechanics involves the assumption that, even though we know that the microstates exist, we are largely ignorant of their actual values. The only information we have about them comes from macroscopic quantities, which are bulk properties such as the total energy, the temperature, the volume, the pressure, or the magnetization. Because of our ignorance we have to treat the microstates in statistical terms. But the knowledge of the macroscopic quantities, along with the laws of physics that the microstates follow, constrain the microstates and allow us to compute relations between macroscopic variables that might otherwise not be obvious. Once the values of the macroscopic variables are fixed there is typically only a subset of microscopic states that are compatible with them, which are called the *accessible states*. The number of accessible states is typically huge, but differences in this number can be very important. In this chapter we will for simplicity assume a discrete set of microstates, but the formalism can be straightforwardly generalized to the continuous case.

The first fundamental assumption of statistical mechanics is that a closed system has an equal a priori probability to be in any of its accessible states. For systems which are not closed, for example because they are in thermal contact or their particle number is not constant, the set of accessible states will be different and appropriate probabilities for them have to be defined. Another way of saying this is that with a real macroscopic physical system in equilibrium we associate an *ensemble* of systems with a characteristic probability distribution over the allowed microscopic states. Tolman [22] clearly described the notion of an ensemble:

> In using ensembles for statistical purposes, however, it is to be noted that there is no need to maintain distinctions between individual systems since we shall be interested merely in the number of systems at any time which

> would be found in the different states that correspond to different regions of phase space. Moreover, it is also to be noted for statistical purposes that we shall wish to use ensembles containing a large enough population of separate members so that the number of systems in such different states can be regarded as changing continuously as we pass from the states lying in one region of the phase space to those in another. Hence, for purpose in view, it is evident that the condition of an ensemble at any time can be regarded as appropriately specified by the density r with which we representative points are distributed over phase space

The second postulate of statistical mechanics, called *ergodicity*, says that time averages correspond to ensemble averages. That is, on one hand we can take the time average by following the deterministic motion of the all the microscopic variables of all the particles making up a system. On the other hand, at a given instant in time we can take an average over all possible accessible states, weighting them by their probability of occurrence. The ergodic hypothesis says that these two averages are the same. We return to the restricted validity of this hypothesis in the section on nonlinear dynamics.

3.2. **A simple model system of magnetic spins.** In the following example we show how it is possible to derive the distribution of microscopic states through the assumption of equipartition and simple counting arguments. The results also illustrates that the distribution over microstates becomes extremely narrow in the thermodynamic (i.e. $N \to \infty$ limit). Consider a system of N magnetic spins that can only take two values $s_j = \pm 1$, corresponding to whether the spin is pointing up or down (often called *Ising spins*). The total number of possible configurations equals 2^N. For convenience assume N is even, and that the spins do not interact. Now put these spins in a magnetic field H (pointing upward), and ask how many configurations of spins are consistent with each possible value of the energy. The energy of each spin is $e_j = \mp \mu H$, and because they do not interact, the total energy of the system is just the sum of the energies of each spin. So for a configuration with k spins pointing up and $N-k$ spins pointing down the total energy can be written as $\varepsilon_m = 2m\mu H$ with $m \equiv (N-2k)/2$ and $-N/2 \leq m \leq N/2$. The value of ε_m is bounded : $-N\mu H \leq \varepsilon_m \leq N\mu H$ and the difference between two adjacent energy levels, corresponding to the flipping of one spin, is $\Delta\varepsilon = 2\mu H$. The number of microscopic configurations with energy ε_m equals

$$g(N,m) = g(N,-m) = \frac{N!}{(\frac{1}{2}N+m)!(\frac{1}{2}N-m)!}. \tag{3.1}$$

We obviously have the quality: $\sum_m g(N,m) = 2^N$. For a thermodynamic system N is really large, so we can approximate the factorials by the Stirling formula

$$N! \cong \sqrt{2\pi N} N^N e^{-N+1/12N+\cdots} \tag{3.2}$$

Some elementary math gives the Gaussian approximation for the binomial distribution for large N:

$$g(N,m) \cong 2^N \left(\frac{2}{\pi N}\right)^{\frac{1}{2}} e^{-2m^2/N} \ . \tag{3.3}$$

We will return to this system later on, but at this point we merely want to show that for large N the distribution indeed becomes a very strongly peaked Gaussian

distribution. The degeneracy of the states around $m = 0$ increases very rapidly, for example $g(50, 0) = 1.264 \times 10^{14}$, but for $N \approx N_A$ one has $g(N_A, 0) \cong 10^{10^{22}}$. Roughly speaking because the width of the distribution grows with $\sqrt{N}$ while the peak height grows as 2^N we see that the distribution strongly narrows with increasing N. If we consider a situation where the total energy of the system is fixed to be $U = m\mu H$ then the a priori probability for finding it with that particular energy is $p_m = 1/g(N, m)$. We will return to this example in the following section to calculate the magnetisation of a spin system in thermal equilibrium.

3.3. **The Maxwell-Boltzmann distribution.** Maxwell and later Boltzmann derived an expression for the probability distribution p_i for a system in thermal equilibrium, i.e. in thermal contact with a heat reservoir kept at a fixed temperature T. For example, an equilibrium distribution function of an ideal gas without external force applied to it does not depend on either position or time, and thus can only depend on the velocities of the individual particles. In general there are interactions between the particles that need to be taken into account. A simplifying assumption, that is well justified by probabilistic calculations, is that processes in which two particles interact at once are much more common than those in which three or more particles interact. If we assume that the velocities of two particles are independent before they interact we can write their joint probability to have velocities v_1 and v_2 as a product of the probability for each particle alone. This implies $p_(v_1, v_2) = p(v_1)p(v_2)$. The same holds after they interact: $p_(v_1', v_2') = p(v_1')p(v_2')$. Clearly in equilibrium, where nothing can depend on time, the probability has to be the same afterward, i.e. $p(v_1, v_2) = p(v_1', v_2')$. How do we connect these conditions before and after the interaction? A crucial observation is that there are conserved quantities that are preserved during the interaction and the equilibrium distribution function can therefore only depend on those. Homogeneity and isotropy of the distribution function selects the (conserved) energy of the particles as the only function on which the distribution depends. The conservation of energy in this situation boils down to the simple statement that $\frac{1}{2}mv_1^2 + \frac{1}{2}mv_2^2 = \frac{1}{2}mv_1'^2 + \frac{1}{2}mv_2'^2$. From these relations Maxwell derived the well known thermal equilibrium velocity distribution:

$$p_0(v) = n\left(\frac{m}{2\pi T}\right)^{3/2} e^{-mv^2/2kT} \tag{3.4}$$

The distribution is a Gaussian. As we saw, to derive it Maxwell had to make a number of assumptions which were plausible but by no means truly fundamental. Boltzmann generalized the result to include the effect of an external conservative force, leading to the replacement of the kinetic energy in (3.4) by the total conserved energy, which includes potential as well as kinetic energy.

Boltzmann's generalization of Maxwell's result makes it clear that the probability distribution p_i for a general system in thermal equilibrium is the famous Maxwell-Boltzmann equilibrium distribution,

$$p_i = e^{-\varepsilon_i/T}/Z\ . \tag{3.5}$$

Z is a normalization factor that ensures the conservation of probability, i.e. $\sum_i p_i = 1$. This implies that

$$Z \equiv \sum_i e^{-\varepsilon_i/T}\ . \tag{3.6}$$

Z is called the *partition function*. The Maxwell-Boltzmann distribution describes the *canonical ensemble*, that is it applies to any situation where a system is in thermal equilibrium and exchanging energy with its environment. This is in contrast to the *microcanonical ensemble*, which applies to isolated systems where the energy is constant, or the *grand canonical ensemble*, which applies to systems that are exchanging both energy and particles with their environment.

To illustrate the power of the Boltzmann-Gibbs distribution let us briefly return to the example of the thermal distribution of Ising spins on a lattice in an external magnetic field. As we pointed out in section (3.2), the energy of a single spin is $\pm\mu H$. According to the Boltzmann-Gibbs distribution, the probabilities of spin up or spin down are

$$p_{\pm} = \frac{e^{\mp \mu H/T}}{Z}. \tag{3.7}$$

The spin antiparallel to the field has lowest energy and therefore is favored. This leads to an average field dependent magnetization m_H (per spin):

$$m_H =< \mu >= \frac{\mu p_+ + (-\mu)p_-}{p_+ + p_-} = \mu \tanh \frac{uH}{T}. \tag{3.8}$$

This example shows how statistical mechanics can be used to establish relations between macroscopic variables that cannot be obtained using thermodynamics alone.

3.4. **Free energy revisited.** In our discussion of thermodynamics we introduced in section 2.2 the concept of the free energy F defined by equation 2.5, and argued that it plays a central role for systems in thermal contact with a heat bath, i.e. systems kept at a fixed temperature T. In the previous section we introduced the concept of the partition function Z defined by equation 3.6. The importance of the partition function Z goes well well beyond its role as a normalization factor, because from the partition function all thermodynamic quantities can be calculated. The free energy is of particular importance, because its functional form leads directly to the definition of entropy in terms of probabilities. At this point it is therefore possible to directly link the thermodynamical quantities to the ones defined in statistical mechanics, by postulating the explicit relation between the free energy and the partition function[4,5]:

$$F = -T \ln Z, \tag{3.9}$$

or alternatively $Z = e^{-F/T}$. From this definition it is possible to calculate basically all thermodynamical quantities for example using the equations (2.9). We now will proceed to derive the appropriate expression for the entropy in statistical mechanics from the previous equation.

---

[4]Once we have identified a certain macroscopic quantity like the free energy with a microscopic expression, then of course the rest follows. Which expression is taken as the starting point for the identification is quite arbitrary. The justification is *a posteriori* in that the well known thermodynamical relations should be recovered, as far as that is actually possible (because the technical problems that arise in realistic calculations for example if the system is not weakly interacting and/or dilute can be quite formidable).

[5]Boltzmann's constant k relates energy to temperature, it's value in conventional units is $1.4 \times 10^{-23} joule/sec$. We have set it equal to unity, which amounts to choosing a convenient unit for energy

3.5. **Gibbs entropy.** From the definition of the free energy in equation (2.5) we obtain that:

$$S = \frac{F - U}{T} \tag{3.10}$$

Now from (3.9) and (3.5) it follows that

$$F = \varepsilon_i + T \ln p_i \ , \tag{3.11}$$

while the equilibrium value for the internal energy is by definition given by

$$U =< \varepsilon >\equiv \sum_i \varepsilon_i \ p_i \tag{3.12}$$

With these expressions for S, F and U, and furthermore recalling that $\sum_i p_i = 1$, we can rewrite the entropy in terms of the probabilities p_i and arrive at the famous expression for the entropy:

$$S = -\sum_i \ p_i \ln p_i \ . \tag{3.13}$$

This expression is usually called the Gibbs entropy[6].

In the special case where the total energy is fixed, the w different (accessible) states all have equal a priori probability $p_i = p = 1/w$. Substitution in the Gibbs formula yields the expression in terms of the number of accessible states, originally due to Boltzmann (and engraved on his tombstone):

$$S = \ln w \tag{3.14}$$

We emphasize the general and crucial feature of this expression, namely that the entropy grows logarithmically with the number of accessible states[7].

At this point we should already mention that for any system with states $\{\psi_i\}$ and a given probability distribution $\{p_i\}$, the Gibbs expression can be considered and turns out to be of great relevance. The formal definition of the amount of information H that can be stored in such a system was introduced in the seminal papers by Shannon [20] in direct analogy to the entropy S,

$$H \equiv -\sum_i p_i \ \log_2 \ p_i. \tag{3.15}$$

As we will extensively discuss in Section 5, this exact quantitative definition of information and its applications transcend the limited origin and scope in conventional thermodynamics and statistical mechanics.

4. Nonlinear dynamics

> The present state of the system of nature is evidently a consequence of what it was in the preceding moment, and if we conceive of an intelligence which at a given instant comprehends all the relations of the entities of this universe, it could state the respective position, motions, and general

[6] In quantum theory this expression is replaced by $S = -Tr\ \rho \ln \rho$ where ρ is the density matrix of the system.

[7] These numbers can be overwhelmingly big. Imagine two macrostates of a system which differ by 1 millicalory at room temperature: the difference in entropy would be given by $\Delta S = -\Delta Q/T = 10^{-3}/293 \approx 10^{-5}$ the the ratio of the number of accessible states would be given by $w_2/w_1 = \exp(\Delta S/k) \approx \exp(10^{18})$, a big number indeed.

> effects of all these entities at any time in the past or future.
>
> Pierre Simon de Laplace (1776)

> A very small cause which escapes our notice determines a considerable effect that we cannot fail to see, and then we say that the effect is due to chance.
>
> Henri Poincaré (1903).

From a naive point of view, statistical mechanics seems to contradict the determinism of Newtonian mechanics. Newton's laws provide a set of differential equations which define a dynamical system ϕ^t, that maps any state $x(0)$ (which is just a vector of measured positions and velocities) into a future state $x(t) = \phi^t(x(0))$. This is completely deterministic. If as Laplace so famously asserted, mechanical objects obey Newton's laws, then why do we need to discuss perfect certainties in statistical terms?

Laplace partially answered his own question:

> ... But ignorance of the different causes involved in the production of events, as well as their complexity, taken together with the imperfection of analysis, prevent our reaching the same certainty [as in astronomy] about the vast majority of phenomena. Thus there are things that are uncertain for us, things more or less probable, and we seek to compensate for the impossibility of knowing them by determining their different degrees of likelihood. So it is that we owe to the weakness of the human mind one of the most delicate and ingenious of mathematical theories, the science of chance or probability.

This answer is only partially right. As Poincaré later showed, even without human uncertainty (or quantum mechanics), when Newton's laws give rise to chaotic dynamics, we inevitably arrive at a probabilistic description of nature. Although Poincaré discovered this in the course of studying the three body problem, stimulated by his interest in the stability of solar system, the answer he found turns out to have relevance for the reconciliation of the deterministic Laplacian universe and statistical mechanics.

4.1. **The ergodic hypothesis.** As we mentioned in the previous section, one of the key foundations in Boltzmann's formulation of statistical mechanics is the *ergodic hypothesis*. Roughly speaking, it is the hypothesis that a given trajectory will eventually find its way through all the accessible microstates of the system, e.g. all those that are compatible with conservation of energy. At equilibrium the average length of time that a trajectory spends in a given region of the state space is proportional to the number of accessible states the region contains. If the ergodic hypothesis is true, then time averages equal ensemble averages, and equipartition is a valid assumption.

The ergodic hypothesis proved to be highly controversial for good reason: It is generally not true. The first numerical experiment ever performed on a computer took place in 1947 at Los Alamos when Fermi, Pasta, and Ulam set out to test the ergodic hypothesis. They simulated a system of masses connected by nonlinear springs. They perturbed one of the masses, expecting that the disturbance would rapidly spread to all the other masses and equilibrate, so that after a long time they would find all the masses shaking more or less randomly. Instead they were quite surprised to discover that the disturbance remained well defined – although

it propagated through the system, it kept its identity, and after a relatively short period of time the system returned very close to its initial state. They had in fact discovered a phenomenon that has come to be called a *soliton*, a localized but very stable travelling disturbance. There are many examples of nonlinear systems that support solitons.

Clearly statistical mechanics a priori does not apply to such systems. Yet once the system is well understood, i.e. the reasons for the extraordinary stability of the collective phenomenon called solitons, it may be possible to incorporate it again in some statistical mechanical treatment. Without going into detail one may for example point out that many of systems which exhibit solitonic excitations, have also very large numbers of (hidden) symmetries leading to many conservation laws beyond the usual ones, and these are of course causing restrictions on the time evolution of the system from a given initial state (think of the role that energy conservation plays). The time evolution of the system then has to obey many hidden constraints leading to a severe restriction on which parts of phase space can be visited.

There are in fact many examples where we know that statistical mechanics works extremely well. There are a few cases, involving idealized models such as the hard sphere gas, where the ergodic hypothesis can actually be proved. But more typically this is not the case. The evidence for statistical mechanics is largely empirical: we know that it works, at least to a very high degree of approximation. But even when it works, there is a considerable body of lore suggesting that the ergodic hypothesis is too strong, and is not strictly true. While trajectories may wander in more or less random fashion around the accessible phase space, they can be blocked from entering certain regions by what are called KAM (Kolomogorov-Arnold-Moser) tori. On KAM tori trajectories make regular motions, and are not chaotic. Furthermore, such trajectories are trapped on the KAM tori, which has a lower dimension than the full accessible phase space. Nonetheless, via mechanisms that are still not well understood, in most circumstances, with sufficiently abitrary nonlinearities, as the number of degrees of freedom increases (e.g. because it contains more particles), KAM tori become less important. The ergodic hypothesis becomes an increasingly better approximation, a typical single trajectory can reach almost all accessible states, and equipartition becomes a good assumption. The necessary and sufficient conditions for this remain an active field of research.

4.2. **Chaos and limits to prediction.** The discovery of chaos makes it clear that Boltzmann's use of probability is even more justified than he realized. When motion is chaotic, two infinitesimally nearby trajectories separate at an exponential rate. This is a geometric property of the nonlinear dynamics. From a linear point of view the dynamics are locally unstable. To make this precise, consider two N dimensional initial conditions $x(0)$ and $x'(0)$ that are initially separated by an infinitesimal vector $\delta x(0) = x(0) - x'(0)$. Providing the dynamical system is differentiable, the separation will grow as

$$\delta x(t) = D\phi^t(x(0))\delta x(0), \tag{4.1}$$

where $D\phi^t(x(0))$ is the derivative of the dynamical system ϕ^t evaluated at the initial condition $x(0)$. For any fixed time t and initial condition $x(0)$, $D\phi^t$ is just an $N \times N$ matrix, and this is just a linear equation. If the motion is chaotic the length of the separation vector δx will grow exponentially with t in at least one

direction, as shown in Figure 3. Nonetheless, at the same time the motion can

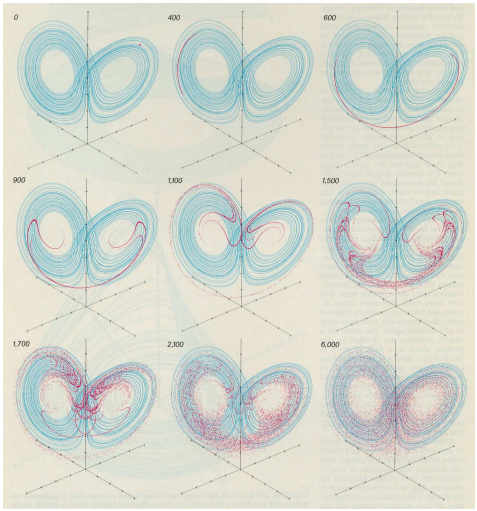

FIGURE 3. Divergence of nearby trajectories is the underlying reason chaos leads to unpredictability. A perfect measurement would correspond to a point in the state space, but any real measurement is inaccurate, generating a cloud of uncertainty. The true state might be anywhere inside the cloud. As shown here for the Lorenz equations (a simple system of three couple nonlinear differential equations [**?**]), the uncertainty of the initial measurement is represented by 10,000 red dots, initially so close together that they are indistinguishable; a single trajectory is shown for reference in light blue. As each point moves under the action of the equations, the cloud is stretched into a long, thin thread, which then folds over onto itself many times, until the points are mixed more or less randomly over the entire attractor. Prediction has now become impossible: the final state can be anywhere on the attractor. For a regular motion, in contrast, all the final states remain close together. We can think about this in information theoretic terms; for a chaotic motion information is initially lost at a linear rate which eventually results in all the information being lost – for a regular motion the information loss is relatively small. The numbers above the illustration are in units of 1/200 of second.

be globally stable, meaning that it remains contained inside a finite volume in the phase space. This is achieved by stretching and folding – the nonlinear dynamics knead the phase space through local stretching and global folding, just like a baker making a loaf of bread. Two trajectories that are initially nearby may later be quite far apart, and still later, may be close together again. This property is called

mixing. More formally, the dynamics are mixing over a given set S and invariant measure[8] μ with support S such that for any subsets A and B

$$\lim_{t\to\infty} \phi^t B \cap A = \mu(A)\mu(B). \tag{4.2}$$

Intuitively, this just means that B is smeared throughout S by the flow, so that the probability of finding a point originating in B inside of A is just the original probability of B, weighted by the probability of A. Geometrically, this happens if and only if the future trajectory of B is finely "mixed" throughout S by the stretching and folding action of ϕ^t.

Mixing implies ergodicity, so any dynamical system that is mixing over S will also be ergodic on S. It only satisfies the ergodic hypothesis, however, if S is the set of accessible states. This need not be the case. Thus, the fact that a system has orbits with chaotic dynamics doesn't mean that it necessarily satisfies the ergodic hypothesis – there may be still be subsets of finite volume in the phase space that are stuck making regular motion on KAM tori.

Nonetheless, chaotic dynamics has strong implications for statistical mechanics. If a dynamical system is ergodic but not mixing[9], by measuring the microstates it is in principle possible to make detailed long range predictions by measuring the position and velocity of all its microstates, as suggested by Laplace. In contrast, if it is mixing then even if we know the initial values of the microstates at a high (but finite) level of precision, all this information is asymptotically lost, and statistical mechanics is unavoidable.

4.3. **Quantifying predictability.** Information theory can be used to quantify predictability. To begin the discussion, consider a measuring instrument with a uniform scale of resolution ϵ. For a ruler, for example, ϵ is the distance between adjacent graduations. If such a measuring instrument is assigned to each of the N real variables in a dynamical system, the graduations of these instruments induce a *partition* Π of the phase space, which is a set of non-overlapping N dimensional cubes, labeled C_i, which we will call the outcomes of the measurement. A measurement determines that the state of the system is in a given cube C_i. If we let transients die out, and restrict our attention to asymptotic motions without external perturbations, let us assume the motion is confined to a set S (which in general depends on the initial condition). We can then compute the asymptotic probability of a given measurement by measuring its frequency of occurrence p_i, and if the motion is ergodic on S, then we know that there exists an invariant measure μ such that $p_i = \mu(C_i)$. To someone who knows the invariant measure μ but knows nothing else about the state of the system, the average information that will be gained in making a measurement is just the entropy

$$I(\epsilon) = -\sum_i p_i \log p_i. \tag{4.3}$$

We are following Shannon in calling this "information" since it represents the element of surprise in making the measurement. The information is written $I(\epsilon)$ to

[8]A measure is invariant over a set S with respect to the dynamics ϕ^t if it satisfies the condition $\mu(A) = \mu(\phi^{-t}(A))$, where A is any subset of S. There can be many invariant measures, but the one that we have in mind throughout is the one corresponding to time averages.

[9]A simple example of a system that is ergodic but not mixing is a dynamical system whose solution is the sum of two sinusoids with irrationally related frequencies.

emphasize its dependence on the scale of resolution of the measurements. This can be used to define a dimension for μ. This is just the asymptotic rate of increase of the information with resolution, i.e.

$$D = \lim_{\epsilon \to 0} \frac{I(\epsilon)}{|\log \epsilon|}. \tag{4.4}$$

This is often called the *information dimension* [**?**]. Note that this reduces to what is commonly called the fractal dimension when p_i is sufficiently smooth, i.e. when $\sum_i p_i \log p_i \approx \log n$, where n is the number of measurement outcomes with nonzero values of p_i.

This notion of dimension can be generalized by using the Renyi entropy R_α

$$R_\alpha = \frac{1}{1-\alpha} \log \sum_i p_i^\alpha \tag{4.5}$$

where $\alpha \geq 0$ and $\alpha \neq 1$. The value for $\alpha = 1$ is defined by taking the limit as $\alpha \to 1$, which reduces to the usual Shannon entropy. By replacing the Shannon entropy by the Renyi entropy it is possible to define a generalized dimension d_α. This contains the information dimension in the special case $\alpha = 1$. This has proved to be very useful in the study of multifractal phenomena (fractals whose scalings are irregular). We will say more about the use of such alternative entropies in the next section.

The discussion so far has concerned the amount of information gained by an observer in making a single, isolated measurement, i.e. the information gained in taking a "snapshot" of a dynamical system. We can alternatively ask how much new information is obtained per unit time by an observer who is watching a movie of a dynamical system. In other words, what is the information acquisition rate of an experimenter who makes a series of measurements to monitor the behavior of a dynamical system? For a regular dynamical system (to be defined more precisely in a moment) new measurements asymptotically provide no further information in the limit $t \to \infty$. But if the dynamical system is chaotic, new measurements are constantly required to update the knowledge of the observer in order to keep the observer's knowledge of the state of the system at the same resolution.

This can be made more precise as follows. Consider a sequence of m measurements $(x_1, x_2, \ldots, x_m) = X_m$, where each measurement corresponds to observing the system in a particular N dimensional cube. Letting $p(X_m)$ be the probability of observing the sequence X_m, the entropy of this sequence of measurements is

$$H_m = -\sum_i p(X_m) \log p(X_m) \tag{4.6}$$

We can then define the information acquisition rate as

$$h = \lim_{m \to \infty} \frac{H_m}{m \Delta t}. \tag{4.7}$$

Δt is the sampling rate for making the measurements. Providing Δt is sufficiently small and other conditions are met, h is equal to the *metric entropy*, also called the *Kolmogorov-Sinai (KS) entropy*[10]. Note that this is not really an entropy, but an

[10]In our discussion of metric entropy we are sweeping many important mathematical formalities under the rug. For example, to make this definition precise we need to take a supremum over all partitions and sampling rates. Also, it is not necessary to make the measurements in N dimensions – there typically exists a one dimensional projection that is sufficient, under an optimal partition

entropy production rate, which (if logs are taken to base 2) has units of bits/second. If $h > 0$ the motion is chaotic, and if $h = 0$ it is regular. Thus, when the system is chaotic, the entropy H_m contained in a sequence of measurements continues to increase even in the limit the sequence becomes very long. In contrast, for a regular motion this reaches a limiting value.

Although we have so far couched the discussion in terms of probabilities, the metric entropy is determined by geometry. The average rates of expansion and contraction in a trajectory of a dynamical system can be characterized by the spectrum of Lyapunov exponents. These are defined in terms of the eigenvalues of $D\phi^t$, the derivative of the dynamical system, as defined in equation 4.1. For a dynamical system in N dimensions, let the N eigenvalues of the matrix $D\phi^t(x(0))$ be $\alpha_i(t)$. Because $D\phi^t$ is a positive definite matrix, the α_i are all positive. The Lyapunov exponents are defined as $\lambda_i = \lim_{t\to\infty} \log \alpha_i(t)/t$. To think about this more geometrically, imagine an infinitesimal ball that has radius $\epsilon(0)$ at time $t = 0$. As this ball evolves under the action of the dynamical system it will distort. Since the ball is infinitesimal, however, it will remain an ellipsoid as it evolves. Let the principal axes of this ellipsoid have length $\epsilon_i(t)$. The spectrum of Lyapunov exponents for a given trajectory passing through the initial ball is

$$\lambda_i = \lim_{t\to\infty} \lim_{\epsilon(0)\to 0} \frac{1}{t} \log \frac{\epsilon_i(t)}{\epsilon(0)}. \tag{4.8}$$

For an N dimensional dynamical system there are N Lyapunov exponents. The positive Lyapunov exponents λ^+ measure the rates of exponential divergence, and the negative ones λ^- the rates of convergence. They are related to the metric entropy by *Pesin's theorem*

$$h = \sum_i \lambda_i^+. \tag{4.9}$$

In other words, the metric entropy is the sum of the positive Lyapunov exponents, and it corresponds to the average exponential rate of expansion in the phase space.

Taken together the metric entropy and information dimension can be used to estimate the length of time that predictions remain valid. The information dimension allows an estimate to be made of the information contained in an initial measurement, and the metric entropy estimates the rate at which this information decays.

As we have already seen, the metric entropy tells us the information gained in each measurement in a series of measurements. But if each measurement is made with the same precision, the information gained must equal the information that would have been lost had the measurement not been made. Thus the metric entropy also quantifies the initial rate at which knowledge of the state of the system is lost after a measurement.

To make this more precise, let $p_{ij}(t)$ be the probability that a measurement at time t has outcome j if a measurement at time 0 had outcome i. (Or another words, given the state was measured in partition element C_i at time 0, what is the probability it will be in partition element C_j at time t?). By definition $p_{ij}(0) = 1$ if $i = j$ and $p_{ij}(0) = 0$ otherwise. With no initial information, the information gained from the measurement is determined solely by the asymptotic measure μ, and is $-\log \mu(C_j)$. In contrast, if C_i is known the information gained on learning outcome j is $-\log p_{ij}(t)$. The extra information using a prediction from the initial

data is the difference of the two or $\log(p_{ij}(t)/\mu(C_j))$. This must be averaged over all possible measurements C_j at time t, and all possible initial measurements C_i. The measurements C_j are weighted by their probability of occurrence $p_{ij}(t)$, and the initial measurements are weighted by $\mu(C_i)$. This gives

$$I(t) = \sum_{i,j} \mu(C_i) p_{ij}(t) \log(\frac{p_{ij}(t)}{\mu(C_j)}). \tag{4.10}$$

It can easily be shown that in the limit where the initial measurements are made arbitrarily precise, $I(t)$ will initially decay at a linear rate, whose slope is equal to the metric entropy. For measurements with signal to noise ratio s, i.e. with $\log s \approx |\log \epsilon|$, $I(0) \approx D_I \log s$. Thus $I(t)$ can be approximated as $I(t) \approx D_I \log s - ht$, and the initial data becomes useless after a characteristic time $\tau = (D_I/h) \log s$.

5. About Entropy

In this section we will discuss various aspects of entropy, its relation with information theory and the sometimes confusing connotations of order, disorder, ignorance and incomplete knowledge. A derivation of the second law using the procedure called *coarse graining* is presented. The extensivity or additivity of entropy is considered in some detail, also when we discuss nonstandard extensions of the definition of entropy.

5.1. **Entropy and information.** The important innovation Shannon made was to show that the relevance of the concept entropy considered as a measure of information, was not restricted to thermodynamics, but could be used in any context where probabilities can be defined. He applied it to problems in communication theory and showed that it can be used to compute a bound on the information transmission rate using an optimal code.

When using entropy in information theory it is common to take logarithms in base two, and to drop the Boltzmann constant[11]. Base two is a natural choice when dealing with binary numbers, and the units of entropy in this case are called *bits*; in contrast, when using the natural logarithm the units are called *nats*, with the conversion that 1 *nat =1.443* bits) . For example a memory consisting of 5 bits (which is the same as a system of 5 Ising spins), has $N = 2^5$ states. Without further restrictions all of these states (messages) have equal probability i.e. $p_i = 1/N$ so that the information content is $H = -N\frac{1}{N} \log_2 \frac{1}{N} = \log_2 2^5 = 5$ *bits*. Similarly, the naive information content of a DNA-molecule with 10 billion base pairs, each of which can be one of four combinations (A-T,C-G,T-A,G-C) can a priori be in any of $4^{10^{10}}$ configurations, the information becomes $H = 2 \times 10^{10}$ *bits*. The logarithmic nature of the definition is almost unavoidable if one wants the additive property of information under the addition of bits. If in the previous spin example we add another string of 3 bits then the total number of states is $N = N_1 N_2 = 2^5 \times 2^3 = 2^8$ from which it also follows that $H = H_1 + H_2 = 8$. If we add extra ab initio correlations or extra constraints the information will decrease, we reduce the number of independent configurations and consequently H will be smaller.

One of the most important results that Shannon gave was to show that the choice of the Gibbs form of entropy is not arbitrary, even in an arbitrary context. Both Shannon and Khinchin [12] proved that if one wants certain conditions to be met

[11] In our convention k=1, so $H = S/\ln 2$.

by the entropy function then this is the unique choice. The fundamental conditions as specified by Khinchin are:

(1) For a given n and $\sum_{i=1}^{n} p_i = 1$, the required function $H(p_1, ...p_n)$ is maximal for all $p_i = 1/n$.
(2) The function should satisfy $H(p_1, ...p_n, 0)H(p_1, ...p_n)$. the inclusion of an impossible event should not change the value of H.
(3) If A and B are two finite sets of events, not necessarily independent, the entropy $H(A, B)$ for the occurrence of joint events A and B shall be the entropy for the set A alone plus the weighted average of the conditional entropy $H_i(B)$ for B given the occurrence of the i^{th} event in A,

$$H(A, B) = H(A) + \sum_i p_i H_i(B) \tag{5.1}$$

where event A_i occurs with probability p_i.

The important result is that given these conditions the function H (eqn. 3.15) is the *unique* solution. Shannon's key insight was that the results of Boltzmann and Gibbs in explaining entropy in terms of statistical mechanics had unintended and profound side-effects, with a broader and more fundamental meaning that transcended their physical origin of entropy. The importance of the abstract conditions formulated by Shannon and Khinchin show the very general context in which the Gibbs-Shannon function is the unique answer. Later on we will pose the question of whether there are situations in physics c.q. information theory, where *not* all three conditions have to be imposed, leading to alternative expressions for the entropy (or information).

5.2. **The entropy as a relative concept.**

> Irreversibility is a consequence of the explicit introduction of ignorance into the fundamental laws.
>
> M. Born

There is a surprising amount of confusion about the interpretation and meaning of the concept of entropy [7, 3]. To what extent is the "entropic principle" just an "anthropocentric principle"? That is, does entropy depend only on our perception, or is it something more fundamental? Is it a subjective attribute in the domain of the observer or is it an intrinsic property of the physical system we study? Let us consider the common definition of entropy as a measure of disorder. This definition can be confusing unless we are careful in spelling out what we mean by order or disorder. For instance, consider the crystallization of a supercooled liquid. Assume a closed system, where no energy is exchanged with the environment. Initially the molecules of the liquid are free to randomly move about, but then (often through the addition of a small perturbation that breaks the symmetry) the liquid crystallizes, the liquid turns into a solid, and the molecules get pinned to the sites of a regular lattice. From one point of view, this a splendid example of creation of order out of chaos. Yet from standard calculations in statistical mechanics we know that the entropy increases during crystallization. This is because what meets the eye is only part of the story. During crystallization entropy is generated in the form of latent heat, which is stored in the vibrational modes of the crystal. Thus, even though in the crystal the individual molecules are constrained to be roughly in a particular location, they randomly vibrate around their lattice sites more energetically than when they were free to wander. From a microscopic point of view there are more accessible states in the crystal than there were in the liquid, and thus the

entropy increases. The thermodynamic entropy is indifferent to whether motions are microscopic or macroscopic – it only counts the number of accessible states and their probabilities.

In contrast, to measure the sense in which the crystal is more orderly, we must measure a different set of probabilities. To do this we need to define probabilities that depend only on the positions of the particles and not on their velocities. To make this even more clear-cut, we can also use a more macroscopic partition, large enough so that the thermal motions of a molecule around its lattice site tend to stay within the same partition element. The entropy associated with this set of probabilities, which we might call the "spatial order entropy", will behave quite differently from the thermodynamic entropy. For the liquid, when every particle is free to move anywhere in the container, the spatial order entropy will be high, essentially at its largest possible value. After the crystallization occurs, in contrast, the spatial order entropy will drop dramatically. Of course, this is *not* the thermodynamic entropy, but rather an entropy that we have designed to quantitatively capture the aspect of the crystalline order that we intuitively perceive.

As we emphasized before, Shannon's great insight was that it is possible to associate an entropy with any set of probabilities. However, the example just given illustrates that when we use entropy in the broader sense of Shannon we must be very careful to specify the context of the problem. Shannon entropy is just a function that reduces a set of probabilities to a number, reflecting how many nonzero possibilities there are as well as the extent to which the set of nonzero probabilities is uniform or concentrated. Within a fixed context, a set of probabilities that is smaller and more concentrated can be interpreted as more "orderly", in the sense that fewer numbers are needed to specify the set of possibilities. Thermodynamics dictates a particular context – we have to measure probabilities in the full state space. Thermodynamic entropy is a special case of Shannon entropy. In the more general context of Shannon, in contrast, we can measure probabilities however we want, depending on what we want to do. But to avoid confusion we must always be careful to keep this context in mind, so that we know what our answer means.

5.3. **The Gibbs paradox.** Another aspect of the debate on the interpretation of entropy goes back to the very origins of thermodynamics as for example in the 'Gibbs paradox'. In its simplest form it concerns the mixing of ideal gases (kept at the same temperature and pressure) after removing a partitioning in a gas container. If the gases on both sides of the partition are different the gases will mix after the partition is removed and because this is an irreversible process the entropy will increase. However, if the gases would have been *identical*, the partition could be placed back, the system would return to the same situation and the change in entropy would jump to zero. Maxwell imagined the situation where the gases were initially supposed to be identical, and only later recognized to be different, in one case one would have no mixing and in the other there would be. In the one case removing the partitioning would be reversible and in the other not, leading to different results for the total entropy. This reasoning led to the uncomfortable conclusion that the notion of irreversibility and entropy would depend on our knowledge. He concluded that the entropy would thus depend on the state of mind of the experimenter and therefore lacked an objective ground. The early accounts of Gibbs and others also echo this subjectivity judgement, talking about "imperfect knowledge

of the system" reflecting "ignorance". On the other hand one could hardly maintain that more knowledge about the system would actually change the course of (physical) events. This fierce debate about whether the statistical nature of our observations was a reflection of the way nature works or just of our imperfect way of dealing with it, our methodology, is still lingering on. The would be "subjectivity" of the notion entropy as a consequence of our incompleteness of knowledge still has its proponents. Others, from a more pragmatic point of view, have made the in our view relevant remark that the notion of entropy depends certainly on what macroscopic constraints are put on the system, and as the formula's for entropy show: if we release some constraints that means that the number of accessible states increases and the entropy will increase and if we add constraints the entropy will be smaller. So entropy does not talk about what we know, but rather about what we (can) impose, what the precise physical context is [11].

In the end it all boils down to the crucial question on what physical characteristics (labels) are assigned to the different degrees of freedom making up the system. This has very much to do with the actual modelling of the system and to what extend the micro states are distinguishable[12]. Macroscopic measurements and manipulations may or may not be able to distinguish these microscopic features in which case the analysis has to be adapted appropriately (the atoms may or may not have color, carry electric charge or magnetic spins).

We conclude that the adequate definition of entropy reflects the objective physical constraints we put on the system, these have nothing to with our lack of knowledge. The 'incompleteness of our knowledge' is an exact and objective reflection of a particular set of macroscopic constraints deliberately imposed on the physical system we want to describe. The system's behavior depends on these constraints as does the calculated and measured entropy.

5.4. **Adding the entropy of subsystems.** We have mentioned the property that entropy is an extensive quantity. Generally speaking the extensivity of entropy means that it has to satisfy the fundamental linear scaling property:

$$S(T, qV, qN) = qS(t, V, N), \quad 0 < q < \infty \tag{5.2}$$

Extensivity translates in additivity of entropies: if we combine two noninteracting (sub)systems (labelled 1 and 2) with entropies S_1 and S_2, then the total number of states will just be the product of those of the individual systems and taking the logarithm, the entropy of the total system S becomes:

$$S = S_1 + S_2. \tag{5.3}$$

Applying this to two spin systems without an external field, the number of states of the combined system is $w = 2^{N_1+N_2}$, we clearly have $w = w_1\, w_2$ and taking the logarithm establishes the additivity of entropy.

However if we allow for a nonzero magnetic field, this result is no longer obvious and requires some more calculation. In Section 3.2 we calculated the number of configurations with a given energy $\varepsilon_k = -k\mu H$ as $g(N, k)$. If we now allow two systems to exchange energy but keep the total energy fixed, then this generates a

[12]The complete resolution of Gibbs' paradox involved the essential quantum mechanical feature of many body systems that identical particles are indistinguishable. For example, there is no such thing as labelling individual electrons, i.e. there is only *one* state describing a system of N electrons different occupying N different one electron states. This introduces a crucial factor N! compared to the "classical" expressions which gave rise to the inconsistencies.

dependence between the two systems and it is not instantly clear what will happen to the total entropy after equilibrium has established itself. Strictly speaking this simple example will show that the extensivity of entropy is not self evident and should be considered as an additional requirement on the theory, which may or may not give an adequate description of physical situation in some given context.

Let the number of spins pointing up in system 1 be k_1 and the number of particles be N_1, and similarly let this be k_2 and N_2 for system 2. The total energy $k = k_1 + k_2$ is conserved, but the energy in either subsystem (k_1 and k_2) is not conserved. The total number of spins, $N = N_1 + N_2$ is fixed, and so are the spins (N_1 and N_2) in either subsystem. Because the systems only interact when the number of up spins in one of them (and hence also the other one) changes, we can write the total number of states for the combined system as

$$g(N,k) = \sum_{k_1} g_1(N_1,k_1) g_2(N_2,k_2), \tag{5.4}$$

where we are taking advantage of the fact that as long as k_1 is fixed, systems one and two are independent. Taking the log of the above formula clearly does not lead to the additivity of entropies because we have to sum over k_1. This little calculation illustrates the remark made before, that since we have relaxed the constraint that each system had a fixed energy to the condition that only the sum of their energies is fixed, the number of accessible states for the total system is increased (the subsystems themselves are no longer closed) and therefore the entropy will increase.

Let us now indicate that extensivity or better additivity of entropy is recovered in the thermodynamic limit in the above example: the conclusion depends on the large sizes of the systems. Let us consider the contributions to the sum in (5.4) as a function of k_1 where the maximal term corresponds say to $k_1 = \hat{k}_1$ We can now write the contribution in the sum where k_1 deviates an amount δ from $\hat{k}_1$ as,

$$\Delta g(N,k) = g_1(N_1, \hat{k}_1 + \delta) g_2(N_2, \hat{k}_2 - \delta) = f(\delta) g_1(N_1, \hat{k}_1) g_2(N_2, \hat{k}_2) \ , \tag{5.5}$$

where the correction factor can be calculated by expanding the g function in the terms around their respective $\hat{k}$ values. It turns out that f is on the order of $f \sim \exp(-2\delta^2)$ so that the contributions to $g(N,k)$ of the nonmaximal terms in the sum (5.4) are exponentially suppressed. Apparently, in the limit that the number of particles goes to infinity we obtain that entropy is again an additive quantity. This exercise shows also that when a system gets large we may replace the averages of a quantity by its value in the most probable configuration, as our intuition would have suggested. From a mathematical point of view this result follows from the fact that the binomial distribution approaches a gaussian for large values of N.

The general statement is as follows. When two subsystems interact, it is certainly possible that the entropy of one decreases at the expense of the other. This can happen, for example because system one does work on system two, so the entropy of system one goes up while that of system two goes down. This is very important for living systems, which collect free energy from their environment and expel heat energy as waste. Nonetheless, the total entropy S of an organism plus its environment still increases, and in fact so would have the sum of the independent entropies of the non interacting subsystems. That is, if at time zero

$$S(0) = S_1(0) + S_2(0) \ , \tag{5.6}$$

then at time t it may be true that

$$S(t) \leq S_1(t) + S_2(t) \ , \tag{5.7}$$

This is due to the fact that only interactions with other parts of the system can lower the entropy of a given subsystem. In such a situation we are of course free to call the difference between the entropy of the individual systems and their joint entropy a *negative* correlation entropy. However, despite this apparent decrease of entropy, we should keep in mind that both the total entropy and the sum of the individual entropies can only increase, i.e.

$$\begin{aligned} S(t) &\geq S(0) \\ S_1(t) + S_2(t) &\geq S_1(0) + S_2(0). \end{aligned} \tag{5.8}$$

The point here is thus that equations (5.7) and (5.8) are not in conflict.

5.5. **The maximal entropy principle of Jaynes.**

> The statistical practice of physicists has tended to lag about 20 years behind current developments in the field of basic probability and statistics.
>
> E.T. Jaynes (1963)

There are two equivalent sets of postulates that can be used as a foundation to derive an equilibrium distribution in statistical mechanics. One is to begin with the hypothesis that equilibrium corresponds to a minimum of the free energy, and the other is that it corresponds to a maximum of the entropy. The latter approach is a relatively modern development. Inspired by Shannon, Jaynes turned the program of statistical mechanics upside down [10]. Starting from a very general set of axioms he showed that under the assumption of equilibrium the Gibbs expression for the entropy is unique. Under Jaynes' approach, any problem in equilibrium statistical mechanics is reduced to finding the set of p_i for which the entropy is maximal, under a set of constraints that specify the macroscopic conditions, which may come from theory or may come directly from observational data [9]. This variational approach removes some of the arbitrariness that was previously present in the foundations of statistical mechanics. The principle of maximum entropy is very simple and has broad application. For example if one maximizes S only under the normalization condition $\sum_i p_i = 1$, then one finds the unique solution that $p_i = 1/N$ with N the total number of states. This is the uniform probability distribution underlying the equipartition principle. Similarly, if we now add the constraint that energy is conserved, i.e. $\sum_i \varepsilon_i p_i = U$, then the unique solution is given by the Maxwell-Boltzmann distribution, eqn (3.5). Choosing the maximum entropy principle as a starting point shows that a clear separation should be made between the strictly physical input of the theory and purely probabilistic arguments.

The *maximal entropy formalism* has a much wider validity than just statistical mechanics. It is widely used for statistical inference in applications such as optimizing data transfer and statistical image improvement. In these contexts it provides a clean answer to the question, "given the constraints I know about in the problem, what is the model that is as random as possible (i.e. minimally biased) subject to these constraints?". A common application is missing data: Suppose one observes a series of points x_i at regular time intervals, but that somehow some of these observations are missing. By treating the known points as constraints, one can ask for the distribution of the value of the missing points that maximizes the entropy, subject to these constraints.

One must always bear in mind, however, that in physics the maximum entropy principle only implies to equilibrium situations, which are only a small subset of the problems in physics. For systems that are not in equilibrium one must take a different approach. Attempts to understand non-equilibrium statistical mechanics have led some researchers to explore the use of alternative notions of entropy, as discussed in Section 5.9.

5.6. **Ockham's razor.**

> Entia non sunt multiplicanda praeter neccessitatem
> (*Entities should not be introduced except when strict necessary)*)
>
> William van Ockham (1285-1347)

Another interesting and important application of information is to "Ockham's razor". This principle states that if two models have equal predictive power, the simplest model is preferable. While this seems like something we can probably all agree on, real problems are typically not this easy, since while model A may have more parameters than model B, it may also fit the data better. In trying to find a model that is most likely to generalize to data that has not been seen, how does one trade off goodness of fit against number of parameters? Building on earlier work of Akaike, Rissenen introduced a framework to think about this problem in information theoretic terms [**?**, **?**]. The basic idea is to treat the deviations between the model and the data as probabilistic events. A model that gives a better fit has less deviation from the data, and hence implies a tighter probability distribution, which translates into a lower entropy. This entropy is then added to the information needed to specify the parameters. A model with lots of parameters will have more information. The best model is the one with the lowest sum. By characterizing the goodness of fit in terms of bits, this approach puts the complexity of the model and the goodness of fit on the same footing, and gives the correct tradeoff between goodness of fit and model complexity, so that the quality of any two models can be compared. This shows how at some level the concept of entropy underlies the whole scientific method, and indeed, our ability to make sense out of the world. Everytime we make a correct judgement about a pattern in the world, we have correctly made a tradeoff between overfitting (fitting every bump even if it is a random variation) and overgeneralization (identifying events that really are different). Even if we do not make the accounting between model complexity perfectly, when we discover and model regularities in the world, we are implicitly relying on a model selection process of this type, and are intuitively making a judgment trading off the information needed to specify the model and the entropy of the fit of the model to the world.

5.7. **Coarse graining and irreversibility.**

> Our aim is not to 'explain irreversibility' but to describe and predict the observable facts. If one succeeds in doing this correctly, from first principles, we will find that philosophical questions about the 'nature of irreversibility' will either have been answered automatically or else will be seen as ill considered and irrelevant.
>
> E.T. Jaynes

The second law of thermodynamics says that for a closed system the entropy will increase until it reaches its equilibrium value. This corresponds to the irreversibility we all know from daily experience. If we put a drop of ink in a glass of water the drop will diffuse through the water and dilute until the ink is uniformly spread

through the water. The increase of entropy is evident in the fact that the ink is initially in a small region (with $p_i = 0$ except for this region, leading to a probability distribution concentrated on a small region of space and hence a low entropy). The system will not return to its original configuration. Although this is not impossible in principle, it is so improbable that it will never be observed[13].

This irreversibility is hard to understand from the microscopic point of view because the microscopic laws of nature that determine the time evolution of any physical system on the fundamental level are all symmetric under time reversal. That is, the microscopic equations of physics, such as $F = ma$, are unchanged under the substitution $t \to -t$. How can irreversibility arise on the macroscopic level if it has no counterpart on the microscopic level?

In fact, if we compute the entropy at a completely microscopic level it is conserved, which seems to violate the second law of thermodynamics. This follows from the fact that momentum is conserved, which implies that volumes in phase space are conserved. This is called Liouville's theorem which in turn implies that the entropy S is conserved. This doesn't depend on the use of continuous variables – it only depends on applying the laws of physics at the microscopic level. It reflects the idea of Laplace, which can be interpreted as a statement that statistical mechanics wouldn't really be necessary if we could only measure and track all the little details.

The ingenious argument that Gibbs used to get around this, and thereby to reconcile statistical mechanics with the second law of thermodynamics, was to introduce the notion of *coarse graining*. This procedure corresponds to a systematic description of what we could call "zooming out". As we have already mentioned, this zooming out involves dividing phase space up in finite regions δ according to a partition Π. Suppose, for example, that at a microscopic level the system can be described by discrete probabilities p_i for each state. Let us start with a closed system in equilibrium, ie we have a uniform distribution over the accessible states. For the Ising system, for example, $p_i = 1/g(N, i)$ is the probability of particular configuration of spins. Now we replace in each little region δ the values of p_i by its average value $\bar{p}_i$ over δ:

$$\bar{p}_i \equiv \frac{1}{\delta} \sum_{i \in \delta} p_i, \tag{5.9}$$

and consider the associated coarse grained entropy

$$\bar{S} \equiv -\sum_i \bar{p}_i \ln \bar{p}_i. \tag{5.10}$$

At time $t = 0$ we have per definition that $S(0) = \bar{S}(0)$. Next we change the situation by removing a constraint of the system so that it is no longer in equilibrium. In other

[13] "Never say never" is a saying of unchallenged wisdom. What we mean here by "never", is inconceivably stronger then "never in a lifetime", or even "never in the lifetime of the universe". Let's make a rough estimate: consider a dilute inert (say helium) gas that fills the left half of a container of volume V. Then we release the gas in the full container and ask what the recurrence time would be, i.e. how long it would take before all particles would be in left half again. A simple argument giving a reasonable estimate, would be as follows. At any given instant the probability for a given particle to be in the left half is 1/2, but since the particles are independent, the probability of $N \sim N_A$ particles to be in the left half is $P = (1/2)^{10^{23}} \approx 10^{(-10^{20})}$. The time estimate is then given $\tau_0/P = 10^{10^{17}}$ *sec*, where τ_0 is some typical time scale in the system (here we took 1 milisec.).

words we enlarge the space of accessible states but choose as an initial condition that the probabilities are zero for the new states. We can then compare the evolution of the fine-grained entropy $S(t)$ and the coarse-grained entropy $\bar{S}(t)$. The evolution of $S(t)$ is governed by the reversible microscopic dynamics and therefore it stays constant, so that $S(t) = S(0)$. To study the evolution of the coarse-grained entropy we can use a few simple mathematical tricks. First, note that because $\bar{p}_i$ is constant over each region with δ elements,

$$\bar{S}(t) = \sum_i \bar{p}_i \ln \bar{p}_i = \sum_i p_i \ln \bar{p}_i \tag{5.11}$$

Then we may write

$$\bar{S}(t) - \bar{S}(0) = \sum_i p_i (\ln p_i - \ln \bar{p}_i) = \sum_i p_i \ln \frac{p_i}{\bar{p}_i} = \sum_i \bar{p}_i (\frac{p_i}{\bar{p}_i} \ln \frac{p_i}{\bar{p}_i}) \tag{5.12}$$

The mathematical inequality $x \ln x \geq (x-1)$ implies that

$$\bar{S}(t) - \bar{S}(0) \geq \sum_i p_i - \sum_i \bar{p}_i = 1 - 1 = 0 \tag{5.13}$$

The equal sign only occurs if $p_i/\bar{p}_i = 1$ throughout, so except for the special case where this is true, this is a strict inequality and the entropy increases. We see how the second law is obtained as a consequence of coarse graining.

The second law describes mathematically the irreversibility we witness when somebody blows smoke in the air. Suppose we make a film of the developing smoke cloud. If we film the movie at an enormous magnification, so that what we see are individual particles whizzing back and forth, it will be impossible to tell which way the movie is running – from a statistical point of view it will look the same whether we run the movie forward or backward. But if we film it at a normal macroscopic scale of resolution, the answer is immediately obvious – the direction of increasing time is clear from the diffusion of the smoke from a well-defined thin stream to a diffuse cloud.

From a philosophical point of view one should ask to what extent coarse graining introduces an element of subjectivity into the theory. One could object that the way we should coarse grain is not decided upon by the physics but rather by the person who performs the calculation. The key point is that, as in so many other situations in physics, we have to use some common sense, and distinguish between observable and unobservable quantities. Entropy does not increase in the highly idealized classical world that Laplace envisioned, as long as we can observe all the microscopic degrees of freedom and there are no chaotic dynamics. However, as soon as we violate these conditions and observe the world at a finite level of resolution (no matter how accurate), chaotic dynamics ensures that we will lose information and entropy will increase. While the coarse graining may be subjective, this is not surprising – measurements are inherently subjective operations. The important point is that in the limit where the coarse graining is sufficiently fine, the results are independent of it. The increase of entropy is an invariant that will be the same for any sensible coarse graining.

A closely related point is that a system is never perfectly closed – there are always small perturbations from the environment that act as a stochastic perturbation of the system, thereby continuously smearing out the actual distribution in phase space and simulating the effect of coarse graining. Coarse graining correctly captures the fact that entropy is a measure of our uncertainty; the fact that

this uncertainty does not exist for regular motions and perfect measurements is not relevant to most physical problems. Surprisingly this point has not been fully understood by many authors, even in contemporary times [18].

5.8. **Coarse graining and renormalization.** In a written natural language not all finite combinations of letters are words, not all finite combinations of words are sentences, and not all finite sequences of sentences make sense. So by identifying what we call meaningful with accessible, what we just said means that compared with arbitrary letter combinations, the entropy of a language is extremely small.

In modern science something similar is happening. We are used to think of the rich diversity of biological, chemical and physical structures as being enormous, yet in terms of the most fundamental degrees of freedom the structures realized in nature only occupy an extremely tiny part of the unconstrained phase space of the fundamental building blocks. The complete hierarchy starting with the most elementary building blocks of matter such as *leptons* and *quarks*, all the way up to living organisms for that matter, is surprisingly restricted. This has to do with the very specific nature of the interactions between these building blocks, of which to our knowledge there are only four. The interactions lead to the formation of a whole hierarchy of bound states, stable composites of increasing complexity. For by now well understood reasons only very particular stable structures are formed. At each new structural level (think of the subsequent level of quarks, of protons and neutrons, of nuclei, of atoms, of molecules etc) there is a more or less autonomous theory describing the physics at that level involving only the relevant degrees of freedom at that scale. At the higher, more macroscopic levels of the hierarchy only the long range interactions (electromagnetism and gravity) play an important role. So moving up one level corresponds indeed to throwing out an enormous part of the phase space available to the fundamental degrees in the absence of interactions.

We may call the structural hierarchy we just described as "coarse graining" at large. In theoretical physics, however, there is a more subtle but also very successful application of coarse graining called *renormalization* [24]. This procedure has also the aim to understand the large scale behavior of certain dynamical systems which have many degrees of freedom. Here one should think for example of quantum field theories of elementary particles or certain many body systems in statistical physics, where one wants to integrate out the effect of small scale thermal or quantum fluctuations and establish how these effect the large scale properties.

The partition function (as dependent on external parameters as temperature an volume but also on external fields etc) is a weighted sum over all micro states. The probability is given by the exponential Boltzmann factor involving the energy function which specifies the microscopic degrees of freedom and all their interaction parameters such as masses and couplings etc. One can systematically study the effect of averaging over the small scale fluctuations (quantum or thermal) after which one obtains an effective theory with (scale dependent) fields and parameters. This is basically what the term renormalization means. Rescaling the theory (i.e. rescaling the relevant physical distance or momentum scale), drives the effective theory along trajectories in the space of parameters i.e. the space of mass and coupling parameters of the effective (rescaled) degrees of freedom. There are many interesting cases where the asymptotic behavior of the theory is then characterized by some (stable) fixed point in the parameter space. This fixed point can be very different in nature, it may describe the ultraviolet (small distance) or infrared (large

scale) asymptotics. We speak of critical behavior, if for example one of the mass parameters in the fixed point goes to zero because in which case the system will exhibit long distance power law correlations. The physics around the fixed point is determined by the eigenvalues of linearized rescalings: these eigenvalues determine the scaling properties (critical dimensions) of the essential correlation functions and these may differ substantially from what one would have expected naively. One speaks of a *nontrivial fixed point* where the fields exhibit *anomalous (scaling) dimensions*.

The renormalization approach just described goes back to the work of Wilson and Kadanoff in the sixties and unifies in an essential way the formalism of quantum field theory and statistical mechanics, it has a wide spectrum of applications. The renormalization program explained why theories which on a small scale could be very complicated with lots of couplings between the different degrees of freedom will migrate through the parameter space where many parameters would renormalize to zero (thus describing interactions which become irrelevant at long distance scales) while others would migrate to certain characteristic asymptotic values. It even explained the universality of in the observed critical behavior, where in very different situations the same scaling dimensions show up. If we want to study macroscopic behavior a lot of the microscopic details turn out to be irrelevant and many very different microscopic theories may on large scales give rise to identical effective theories. That was the main lesson taught by the renormalization program in quantum field theory (of elementary particles but also of condensed matter systems) and statistical mechanics. It is a subtle formalism by which one in many cases is able to separate the relevant from the irrelevant microscopic information.

5.9. **Beyond the Boltzmann, Gibbs and Shannon entropy: the Tsallis entropy .**

> The equation $S = k \log W + const$ appears without an elementary theory - or however one wants to say it - devoid of any meaning from a phenomenological point of view.
>
> A. Einstein (1910)

As we have already stressed, the definition of entropy as $-\sum_i p_i \log p_i$ and the associated exponential distribution of states apply only for systems in equilibrium. Similarly, the requirements for an entropy function as laid out by Shannon and Khinchin are not the only possibilities. By modifying these assumptions there are other entropies that are useful. We have already mentioned the Renyi entropy, which has proved to be valuable to describe multi-fractals. Another context where this has been shown to be true concerns power laws. Power laws are ubiquitous in both natural and social systems. A power law[14] is something that behaves for large x as $f(x) \sim x^{-\alpha}$, with $\alpha > 0$. Power law probability distributions decay much more slowly for large values of x than exponentials, and as a result have very different (and less well-behaved) statistical properties. Power law distributions are observed in phenomena as diverse as the energy of cosmic rays, fluid turbulence, earthquakes, flood levels of rivers, the size of insurance claims, price fluctuations, the distribution of individual wealth, city size, firm size, government project cost overruns, film sales, and word usage frequencies. Many different detailed models

[14]It is also possible to have a power law at zero or any other limit, and to have $\alpha < 0$, but for our purposes here most of the examples of interest involve the limit $x \to \infty$ and positive α.

can produce power laws, but so far there is no unifying theory, and it is not yet clear whether any such unifying theory is even possible. It is clear that power laws can't be explained by equilibrium statistical mechanics, where the resulting distributions are always exponential. In fact, a common properties of all the physical systems that are known to have power laws, and the models that purport to explain them, is that they are nonequilibrium systems. The ubiquity of power laws suggest that there might be a nonequilibrium generalization of statistical mechanics for which they are the standard probability distribution in the same way that the exponential is the standard in equilibrium systems.

From simulations of model systems with long-range interactions (such as stars in a galaxy) or systems that remain for long periods of time at the "edge of chaos", there is mounting evidence that such systems can get stuck in nonequilibrium metastable states with power law probability distributions of their states for very long periods of time before they finally relax to equilibrium. Alternatively, power laws also occur in many systems that are driven away from equilibrium, and will never relax to equilibrium.

From a purely statistical point of view it is interesting to ask what type of entropy functions are allowed if we alter the last Khinchin postulate, which is the least obvious. Which entropy functions satisfy the remaining two conditions, and some sensible alternative for the third? It turns out that there is at least one interesting class of solutions called q-entropies introduced in 1988 by Tsallis [23, 5]. The parameter q is usually referred to as the *bias* or *correlation* parameter. For $q \neq 1$ the expression for the q-entropy S_q is:

$$S_q[p] \equiv \frac{1 - \sum_i p_i^q}{q - 1} \tag{5.14}$$

For $q = 1$, S_q reduces to the standard Gibbs entropy by taking the limit as $q \to 1$. Following the Jaynes' approach to statistical mechanics, one can maximize this entropy function under suitable constraints to obtain distribution functions that exhibit power law behavior for $q \neq 1$. These functions are called q-exponentials and are defined as:

$$e_q(x) \equiv \begin{cases} [1 + (1-q)x]^{1/(1-q)} & (1 + (1-q)x > 0) \\ 0 & (1 + (1-q)x < 0). \end{cases} \tag{5.15}$$

An important property of q-exponentials is that for $q < 1$ and $x \gg 0$ or $q > 1$ and $x \ll 0$ 1 they have a power law decay. The inverse of the q-exponential is the $ln_q(x)$ function which is given by:

$$\ln_q \equiv \frac{x^{1-q} - 1}{1 - q}. \tag{5.16}$$

The q-exponential can also be obtained as the solution of the equation

$$\frac{dx}{dt} = x^q. \tag{5.17}$$

This is the typical behavior for a dynamical system at the edge of linear stability, where the first term in its Taylor series vanishes. This gives some alternative insight into one possible reason why such solutions may be prevalent. Other typical situations involve long range interactions (such as the gravitational interactions between stars in galaxy formation) or nonlinear generalizations of the central limit theorem.

At first sight a problem with q-entropies is that for $q \neq 1$ they are not additive for systems that are statistically independent. In fact the following equality holds:

$$S_q[p^{(1)}p^{(2)}] = S_q[p^{(1)}] + S_q[p^{(2)}] + (1-q)S_q[p^{(1)}]S_q[p^{(2)}] \tag{5.18}$$

with the corresponding product rule for the q-exponentials:

$$e_q(x)e_q(x) = e_q(x + y + (1-q)xy) \tag{5.19}$$

This is why the q-entropy is often referred to as non-extensive entropy. However, this is probably a blessing in disguise, namely, if the appropriate type of scale invariant correlations between subsystems are typical, then the q-entropies for $q \neq 1$ are strictly additive. The question remains how generic such correlations are and which physical systems exhibit them, though at this point quite a lot of empirical evidence is accumulating to suggest that such functions are at least a good approximation in many situations.

This alternative statistical mechanical theory involves another convenient definition which makes the whole formalism look like the "old" one. Motivated by the fact that the Tsallis entropy weights all probabilities according to p_i^q, it is possible to define an "escort" distribution $P_i^{(q)}$ [1]

$$P_i^{(q)} \equiv \frac{(p_i)^q}{\sum_j (p_j)q}, \tag{5.20}$$

as introduced by Beck. One can then define the corresponding expectation values of a variable A in terms of the escort distribution as

$$< A >_q = \sum_i P_i^{(q)} A_i. \tag{5.21}$$

With these definitions the whole formalism runs parallel to the Boltzmann-Gibbs program.

The framework described above is still in development and may well turn out to be relevant to 'statistical mechanics' not only in nonequilibrium physics, but also in quite different arenas, such as economics.

6. Black Holes: a space time information paradox

In this section we make a small excursion into the realm of curved space-time. Einstein's theory of general relativity unified the concepts of space-time and gravity such that the gravitational force manifests itself through the curvature of space-time. It is the curvature of space-time that determines how matter and radiation propagate, while at the same time it is the matter and radiation content of space-time that determines how space-time is curved. Under general relativity, gravity is no longer an external force, but instead is completely taken into account by the curvature of space-time, and the fact that matter and radiation move along geodesics (shortest paths) in space-time.

A totally unexpected result from general relativity was the prediction of new mysterious objects called black holes. A black hole is what is left after a very massive star has burnt all of its nuclear fuel and subsequently collapses under its own gravitational pull into an ultra compact object. The space-time curvature is so strong that not even light can escape - hence the term "black hole". The escape velocity from a black hole is larger then the speed of light, which means that - at least classically - no information from inside the black hole can ever reach us. The

physical size of a black hole is defined by its event horizon, which is an imaginary sphere centered on the black hole with a radius (called the Schwarzchild radius)

$$R_s = 2G_N M \ , \tag{6.1}$$

where G_N is Newton's gravitational constant and M is its mass. For a black hole with the mass of the sun this would correspond to $R_S = 3km$, and for the earth only $R_S = 1cm$! From a classic point of view, nothing can ever escape from inside the event horizon. The only measurable quantities of a black hole are its mass, its charge and its angular momentum. From a purely classical point of view, all other information that falls into the black hole, such as the shape of table, a person's face, or the Encyclopedia Britannica is lost, and no trace is left except insofar the addition of these to the black hole influence the mass, charge, and angular momentum.

Quantum mechanics makes a somewhat different prediction. This is because the quantum mechanical states are described by wave functions which evolve in time via dynamics that preserves volume in phase space (a so called unitary time evolution of the wave function). Under essentially parallel arguments to Liouville's theory mentioned earlier, this means that quantum mechanical information is preserved. How can this be compatible with the fact that nothing can escape from the black hole? This gave rise to fundamental debate in physics between the two principle theories of nature: the theory of relativity describing space-time and the theory of quantum mechanics describing matter and radiation. Would the geometry of Einstein's theory of relativity overthrow quantum theory, or visa versa?

The first encounter between these two theories was due to Steven Hawking, who in 1975 showed that if we take quantum theory into account black holes aren't black at all! Instead, he showed that they would emit black body thermal radiation (just like a black stove) at a specific temperature, called the Hawking temperature, given by

$$T_H \equiv \frac{\hbar c}{4\pi R_S} = \frac{\hbar c}{8\pi G_N M} \tag{6.2}$$

This temperature of the black hole is inversely proportional to its mass, which means that a black hole radiates more energy as it becomes lighter. In other words, a black hole will radiate and lose mass at an ever-increasing rate until it finally explodes [15]. Bekenstein and Hawking furthermore showed that it is possible to assign an entropy to a black hole through thermodynamic arguments. This entropy is proportional to the area A of the event horizon, which is $A = 4\pi(R_S)^2)$:

$$S = \frac{A}{4G_N h}. \tag{6.3}$$

This gives rise to a controversy about information: If we throw

an encyclopedia into a black hole, containing say, a gigabit of information, what happens to it? it disappears in the black hole, gets chewed up, and then as the black hole evaporates, the material substance of the encyclopedia provides some mass that may be later emitted as radiation. But does this radiation contain any

[15]We think of blackholes are very massive objects like collapsed stars. The lifetime of such very heavy objects is enormous, so the radiation process is exponentially slow meaning that the lifetime of such a black hole would be $\geq 10_40$ years. Theoretical physicists consider also microscopic black holes and that is where the information paradox we are discussing really leads to a principal contradiction.

trace of the what was originally in the encyclopedia? Could a clever detective making careful measurements with very fancy equipment ever recover it? If not, it would seem that the information is lost, and the laws of quantum mechanics are violated. What cherished principles must be given up to resolve this? One aspect that comes to mind is the striking parallel with thermodynamics, strongly suggesting that one needs to look for an underlying statistical mechanics to explain the formula for black hole entropy. But what are the corresponding microscopic degrees of freedom that exist inside the black hole?

This leads to yet another peculiarity of black holes. As we explained before, the entropy of systems that are not strongly coupled is an extensive property, which is proportional to volume. Here the entropy is indeed extensive, but it is proportional to the area of the event horizon rather than the volume of the black hole. This dimensional reduction of the number of degrees of freedom is highly suggestive that all the physics of a black hole takes place at its horizon, an idea that is called the "holographic principle". Woth a perfect hologram it would be hard two distinguish the three dimensional object from the essentially two dimensional surface of the hologram.

Resolving the clash between the quantum theory of matter and the general relativity of space-time is one of the main motivations for the great effort to search for a theory that overarches all of fundamental physics. At this moment the main line of attack is superstring theory, which is a quantum theory in which both matter and space time are a manifestation of extremely tiny strings ($l = 10^{-35}m$). This theory incorporates microscopic degrees of freedom that might provide a statistical mechanical account of the entropy of black holes. In 1996 Andrew Strominger and Cumrun Vafa managed to calculate the Bekenstein-Hawking entropy for certain simple black holes in terms of microscopic strings and related concepts, by counting the number of accessible quantum states. The answer they found is that for the exterior observer information is preserved on the surface of the horizon. This means nothing less than that the would be lost information can in principle be recovered fully consistent with the postulates of quantum theory. The solution formed an example of the holographic principle and the communis opinio - at least for the moment - is that the principles of quantum theory have successfully passed a severe test[16].

7. Information in design and engineering

Also in many engineering disciplines the notion of information flourishes. As an example we briefly discuss the use of the information concept in the formal (even axiomatic) approach to design problems [21]. Starting point in a rational design process is a chain of functional relationships or mappings, for example one may break up the trajectory from customer demands to final product design into different rather independent subprocesses

(1) Customer demands ⇒ Functional requirements
(2) Functional requirements ⇒ Physical domain design parameters
(3) Physical domain ⇒ Production domain

[16]A long standing bet between Steven Hawking and John Presskil of Caltech was settled in 2004 when Hawking officially declared defeat

Each step has its own functional relationship which ought to be optimized by the 'designer'. At each level the variables on the left of the arrow are expressed as functions of the variables on the right hand side and the challenge is to meet the specifications on the left hand side by adjusting the domain and parameter values on the right hand side. In other words a generic step in the design hierarchy can now be described as:

- an initial set of functional requirements $F_1, F_2...$, which we can collect into a vector $\overrightarrow{F}$ with corresponding components $\{F_i \mid i = 1, ..., N\}$. These requirements should be independent.
- there will be a set of initial design parameters $D_1, D_2...$ collected in a vector $\overrightarrow{D}$ with components $\{D_j \mid j = 1, ..., M\}$

The goal of the design process is clearly to find an optimal matching between the functional requirements and some point in design parameter space. A priori we can make the following general remarks. The first is usually referred to as the first axiom of design theory, which is the statement that functional requirements should be independent. In mathematical terms they span an orthogonal basis for the vector space in which $\overrightarrow{F}$ lives. The second remark concerns the dimensionalities of the various spaces: let us consider the dimension M of the design parameter space and the dimension N of the requirement space. Clearly, if $M < N$ it will in general be hard to find a satisfactory solution to all requirements, while if $M > N$, then generically the opposite will be the case, there is a redundancy in the design, which usually means that we are nor heading for the most efficient solution to our design problem. A simpler design should be possible which meets all requirements. Indeed, the optimal situation appears to occur if $M = N$. Therefore the designer will strive for a situation where $N = M$.

Let us now focus on a single step of the design chain mentioned before. So, a set of functional requirements will be systematically coupled to the set of design parameters. In general we may think of the F_i as functions of the $\{D_j\}$in other words the functional requirements $\overrightarrow{F}$ form a vector field over the design parameter space. The designer chooses an initial point in design space $D_i = D_i^0 (i = 1, ...N)$ and then will study the linearized problem around that point. Indeed this choice of D_i^0 underscores the importance of the preferences and the qualities of the designer or the team of designers. This dependence can be expressed in the following way:

$$dF_i = \sum_j \left[\frac{\partial F_i}{\partial D_j}\right]_{\overrightarrow{D}=\overrightarrow{D}^0} dD_j \tag{7.1}$$

This design equation can now be used to calculate the variation in $\Delta\overrightarrow{F}$ also called *system range* corresponding to the given design parameter tolerances $\Delta\overrightarrow{D}$. Clearly $\Delta\overrightarrow{F}$ is here defined as a quantity derived from the design parameter tolerances. On the other hand the "client" will from the onset have specified the *design range* in the functional requirements as $\delta\overrightarrow{F}$, and the designer clearly wants to maximize the overlap between $\overrightarrow{F}^0 + \delta\overrightarrow{F}^0$ and $\overrightarrow{F}(\overrightarrow{D}^0) + \Delta\overrightarrow{F}$ at minimal cost. In principle we can introduce a system probability density function $p_s(F_1, F_2, ..., F_N)$ on the system range. This probability density is measure for the relative success of the design depending on the specific values of the design parameters. The overlap between the system probability density function and the design range is called the *common range*. Integrating the system probability density over the common (or

design range) produces a total probability which is a measure for the success of the design.

$$p_d = \int_{design\ range} p_s(F_1, F_2, ..., F_N)\ dF_1 dF_2 ... dF_N \tag{7.2}$$

Here we can introduce the notion of information again as minus the log of the total probability of the design.

$$I_d = -\log_2\ p_d \tag{7.3}$$

The meaning of this definition is best characterized by the notion of *information need*, this follows from the following considerations. Imagine the overlap range to be equal to the design range then one can obviously meet the requirement without additional information (or effort), if one makes the tolerance smaller, one clearly needs additional information (or 'skill') to meet the requirements. So from this point of view it is clear that in the design process we want to minimize the information content of the process. So information in context of design is the measure of knowledge required to satisfy a given functional requirements F_i at a given level in the design chain/hierarchy. To have the largest probability for success (with the least effort i.e. costs) we therefore have to *minimize the information associated with the design*. This is the second axiom of design theory.

The question of how - if at all - this can be done is crucially dependent on the precise form of the design equation (7.1), that is on the structure of the matrix

$$A_{ij} = \left[\frac{\partial F_i}{\partial D_j}\right]_{\vec{D}=\vec{D}^0} \tag{7.4}$$

We conclude this section by making the following observations: Firstly we note that we are free in choosing the order in which we put the the requirements in the vector $\overrightarrow{dF}$ and similarly for the entries of the vector $\overrightarrow{dD}$. The fact that we may permute the elements of both vectors independently amounts to the possibility of introducing two permutation matrices P and P' such that $\tilde{F} \equiv P\overrightarrow{F}$ and $\tilde{D} \equiv P'\overrightarrow{D}$. The problem is then equivalent to the problem $d\tilde{F} = \tilde{A}d\tilde{D}$, where $\tilde{A}$ is defined as $\tilde{A} = PA(P')^{-1}$. At this point the following remark should be made: one may want to keep the order of the elements in the vectors for example because the order indicates the relative importance of the requirements and of the design parameters. This does not affect the following discussion in any essential way.

In design theory the following important distinctions are made regarding the structure of the matrix $\tilde{A}$, i.e. after suitable permutations have been made (where from now on we also assume $N = M$) :

- If the matrix $\tilde{A}$ is diagonal, the problem trivializes because the system is uncoupled, which means that each F_i is only dependent on a single design parameter. So, either the requirement can be met by an adjustment of each parameter or not, in the latter case one has to go back one step in the process.
- If the matrix $\tilde{A}$ is triangular then the problem is decoupled and there is a systematic approach to the problem.
- If the matrix $\tilde{A}$ is "randomly filled" then the system is essentially coupled and it is very hard to develop a systematic design strategy.

So it is clear that in the nonlinear space we have to search for a point where the problem becomes decoupled. Then we should continue from there. Let us look at the optimal strategy for a simple example of a decoupled system:

$$\begin{pmatrix} F_1 \\ F_2 \\ F_3 \end{pmatrix} = \begin{bmatrix} a_1 & 0 & 0 \\ a_2 & b_2 & 0 \\ a_3 & b_3 & c_3 \end{bmatrix} \begin{pmatrix} D_1 \\ D_2 \\ D_3 \end{pmatrix} \tag{7.5}$$

Clearly We first choose D_1 to satisfy F_1, having fixed D_1 we subsequently choose D_2 to satisfy the second member of the vector equation, in this context written appropriately as

$$b_2 D_2 = F_2 - a_2 D_1 \ .$$

This procedure should then be continued. In other words the structure of the matrix tells us what the strategy towards a successful design, is.

8. Conclusion

In this chapter we have briefly reviewed the extensive interface between physics and information theory. This lead us naturally back to some of the long standing debates in physics on the nature of the concept of entropy in its many guises. After introducing the key elements of classical thermodynamics, statistical mechanics and nonlinear dynamics, we touched on the foundational work of Shannon on information theory and the powerful impact it had in turn on physics as manifested in the work of Jaynes. We spent an important part of the chapter on several, rather philosophical aspects of entropy and tried to give an up to date account of how we think about this in various scientific contexts. We concluded with some excursions on special topics such as alternative expressions for the entropy applicable in certain nonequilibrium situations, the black hole entropy and its paradoxes and the use of information theory in design theory.

It is clear that the notions of entropy and information are alive and well, and there are many new areas in science where new uses of them are explored with remarkable success.

References

1. C. Beck, Phys.Rev.Lett. **87** (2001), 180601.
2. L. Boltzmann, *Vorlesungen über Gastheorie*, university of california press, berkeley, 1964 ed., translated by S.G. Brush, vol. I and II, J.A. Barth, Leipzig, *Lectures on Gas Theory*, 1896-1898.
3. K.G. Denbigh and J.S. Denbigh, *Entropy in relation to incomplete knowledge*, Cambridge University Press, 1985.
4. R.P. Feynman, *There is plenty of room at the bottom*, Engineering and Science, Caltech (February, 1959).
5. M. Gell-Mann, *Nonextensive entropy: Interdisciplinary applications*, Proceedings Volume in the Santa Fe Institute Studies in the Sciences of Complexity, Oxford University Press, 2004.
6. J.W. Gibbs, *Elementary principles in statistical physics*, Yale University Press, 1902.
7. Y.M. Guttman, *The concept of probability in statistical physics*, Cambridge University Press, 1999.
8. K. Huang, *Statistical mechanics*, Wiley and Sons, 1987.
9. E.T. Jaynes, *Information theory and statistical mechanics*, in Statistical Physics (K. Ford ed.), Benjamin, New York, 1963.

10. ———, *Papers on probability, statistics and statistical physics*, (R.D. Rosenkranz ed.), Reidel, Dordrecht, 1983.
11. ———, *The Gibbs paradox*, Maximum Entropy and Baysean Methods (G. Erickson, P. Neudorfer and C.R. Smith eds.), Kluwer, Dordrecht, 1992.
12. A.I. Khinchin, *Mathematical foundations of statistical mechanics*, Dover,New York, 1949.
13. C. Kittel, *Elementary statistical physics*, Wiley and Sons, 1966.
14. R. Landauer, *Irreversibility and heat generation in the computing process*, IBM Journal of Research and Development **5** (1961), 183–191.
15. ———, *Information is physical*, Physics Today **44** (1991), 23–29.
16. E.M. Lifschitz and L.D. Landau, *Statistical physics*, Course of Theoretical Physics, vol. 5, Butterworth-Heinemann, 1980.
17. J.C. Maxwell, *Theory of heat*, D. Appleton & Co, New York, 1872.
18. I. Prigogine, *From being to becoming: Time and complexity in the physical sciences*, Freeman&Co, 1981.
19. F. Reif, *Fundamentals of statistical and thermal physics*, McGraw-Hill, 1965.
20. C.E. Shannon, *A mathematical theory of communication*, The Bell System Technical Journal **27** (1948), 379–423,623–656.
21. N.P. Suh, *Principles of design*, Oxford University Press, 1990.
22. R.C. Tolman, *Principles of statistical physics*, Clarendon, Oxford, UK, 1938.
23. C. Tsallis, J.Stat.Phys. **52** (1988), 479.
24. J. Zinn-Justin, *Quantum field theory and critical phenomena*, Clarendon Press, Oxford, UK, 1989.

E-mail address: bais@science.uva.nl , jdf@santafe.edu

Santa Fe Institute