On the Relation of the Laws of Thermodynamics to Statistical Mechanics

Wayne C. Myrvold
Department of Philosophy
The University of Western Ontario
wmyrvold@uwo.ca

July 23, 2021

Abstract

Much of the philosophical literature on the relations between thermodynamics and statistical mechanics has to do with the process of relaxation to equilibrium. There has been comparatively little discussion of how to obtain what have traditionally been recognized as laws of thermodynamics, the zeroth, first, and second laws, from statistical mechanics. This note is about how to obtain analogues of those laws as theorems of statistical mechanics. The difference between the zeroth and second laws of thermodynamics and their statistical mechanical analogues is that the statistical mechanical laws are probabilistically qualified; what the thermodynamical laws say will happen, their statistical mechanical analogues say will *probably* happen. For this reason, it is entirely appropriate — indeed, virtually inevitable — for the quantities that are statistical mechanical analogues of temperature and entropy to be attributes of probability distributions. I close with some remarks about the relations between so-called "Gibbsian" and "Boltzmannian" methods in statistical mechanics.

Contents

1 Introduction

This note is about the relations between statistical mechanics and the laws of thermodynamics, and how analogues of the laws of the thermodynamics may be obtained from statistical mechanics. None of the results to be discussed are new, but it seems that how it is that one obtains analogues of the laws of thermodynamics, and what these analogues are, are matters that are not as widely appreciated in philosophical discussions of the relations between the two theories as they should be. Furthermore, textbooks of statistical mechanics do not always lay things out as clearly as could be.

I speak of obtaining *analogues* of the laws of thermodynamics from statistical mechanics. The reason for this is that (as Maxwell was the first to clearly articulate), the kinetic theory of heat entails that the laws of thermodynamics, as originally conceived, cannot be strictly true. Thermodynamic states are characterized by a small number of macroscopically ascertainable parameters, and, once it is accepted that the systems treated of in thermodynamics consist of a vast number of molecules, it must be admitted that the parameters used to characterize a thermodynamic state determine at best a minuscule fraction of the full degrees of freedom of a system, and that a specification of these parameters is not even remotely close to being sufficient for determining, on the basis of the dynamical laws of the system alone, the subsequent behaviour of the system. Moreover, there is no feasible way to prepare a system in such a way as to rule out microstates that will lead to deviations from the behaviour prescribed by the laws of thermodynamics. The founders of statistical mechanics concluded that macroscopically observable departures from the laws of thermodynamics are not to be thought *impossible*, but, at best, extremely *improbable* within the regime in which thermodynamics is empirically successful.

The statistical mechanical analogues of the laws of thermodynamics must, therefore, be probabilistically qualified statements. This means that we will have to make sense of probabilistic talk. How to make sense of probabilities in physics is, of course, the subject of substantial philosophical literature. We will not delve into this topic here; as long as sense can be made of probabilistically qualified assertions, this will serve our purpose. A *caveat* is in order, however. There have been attempts to construe probabilistic assertions about a physical system as involving implicit reference to some actual or hypothetical ensemble of systems, a view that is known as *frequentism*, and which was on the rise at the time that Boltzmann introduced the use of *ensembles* of similarly prepared systems into the literature on statistical mechanics. Though a frequentist view of probability continues to be espoused by many physicists, in the philosophical literature on interpretations of probability there seems to be, more or less, a consensus that frequentism is not a viable view of the meaning of probabilistic statements (see La Caze 2016 and Myrvold 2021a, §3.2 for discussion). Though probabilistic judgments may be informed by multiple experiments performed on similarly prepared systems, we should not conflate the *evidence* for statements of probability with their *meaning*.[1]

For the purposes of this note, it won't matter whether the mechanics underlying our statistical mechanics is classical or quantum.[2] There's a reason for this. We'll be concerned, not with the full set of degrees of freedom of a physical system, but with a restricted set of macrovariables (and indeed, we'll be mostly concerned with one macrovariable, energy, and its transfer between systems). The restriction of a pure quantum state to a set of commuting observables is (unless it's joint eigenstate of those observables) a mixed state, and, as far as those observables are concerned, acts like a classical probability distribution.[3]

The fact that, on the kinetic theory of heat, the laws of thermodynamics cannot be strictly true, has consequences for the concepts of thermodynamics that are not often emphasized. The zeroth law of thermodynamics is a *presupposition* of the definition of temperature; that the law hold is a necessary condition for the existence of a function T of the physical states of systems, equal values of which indicate that there will be no net heat exchange between systems placed in thermal contact. In a regime in which deviations from the zeroth law cannot be neglected, there *can be no such function*. The second law of thermodynamics is a presupposition of the definition of thermodynamic entropy; the law is a necessary condition for the existence of a function S of the physical state of the system that satisfies the definition of

[1] For discussion of what sense *can* be made of probabilistic assertions in the context of statistical mechanics, see Myrvold (2016) and Ch. 8 of Myrvold (2021a).

[2] This is not to say that the differences between classical and quantum mechanics don't matter at all to thermodynamics. It's just that the derivations of the laws of thermodynamics that are the subject of this note are valid in both classical and quantum mechanics.

[3] Can all probabilities in statistical mechanics, even classical statistical mechanics, be thought of as quantum probabilities? This is an interesting question. An affirmative answer has been argued for by David Wallace (2016). For the purposes of this note, we can remain non-committal on this question.

thermodynamic entropy. In a regime in which deviations from the second law cannot be neglected, there can be no such function. One might, nevertheless, seek to define statistical mechanical analogues of temperature and entropy, that play roles, in the statistical mechanical analogues of the laws of thermodynamics, that parallel the roles played in the laws of thermodynamics by their thermodynamic counterparts.

We should expect to obtain from statistical mechanics analogues of the laws of thermodynamics that, in an appropriate regime, yield assertions that closely approximate the laws of thermodynamics.[4] The fact that the statistical-mechanical analogues of the laws of thermodynamics are to be couched in probabilistic terms, whereas the laws of thermodynamics are not, is a clue as to how this will go. It is a familiar fact that, under suitable conditions, usually ones that involve some version of the weak law of large numbers, probability distributions can yield probabilities that are close to certainty. One way for this to happen is for the probability distribution for some quantity to be tightly focussed on the expectation value of that quantity, so that non-negligible deviations of the quantity from its expectation value have negligible probability. When this holds, we may treat the quantity as if it is certain to be equal to its expectation value, and use the actual value and the expectation value interchangeably in our calculations.

Call the regime in which assertions can be made, on the basis of statistical mechanics, about the value of the value of some quantity, with a degree of probability that departs only negligibly from complete certainty, the *quasi-deterministic regime*. It is in the quasi-deterministic regime that we will recover the laws of thermodynamics.

As the zeroth law and second law are presuppositions of the definitions of temperature and thermodynamic entropy, respectively, these quantities will be undefined outside the quasi-deterministic regime. Given probabilistic analogues of the laws of thermodynamics, there may be quantities, well-defined even outside the quasi-deterministic regime, that play roles in the probabilistic analogues of the laws of thermodynamics that are analogous to the roles of temperature and entropy in the original versions of those laws, and which, in the quasi-deterministic regime, have values that with near certainty are near to the values of their thermodynamic counterparts. In

[4] It is common to express the relations between theories in terms of limiting relations. I have expressed the relation, not in terms of limits, but in terms of approximation within a certain regime, because it is not essential that one be able to obtain one theory from the other by taking a limit as some parameter or parameters approach limiting values.

section 3, below, probabilistic analogues of the laws of thermodynamics will be presented, which are obtained from the original versions of the laws of thermodynamics by replacing talk of energy exchanges, unqualified by probabilistic considerations, with talk of *expectation values* of energy exchanges, and which invoke quantities that bear the relations to expectation values of energy exchanges that thermodynamic temperature and entropy bear to actual energy exchanges. Inevitably, since expectation values are attributes of probability distributions, the statistical mechanical analogues of temperature and entropy invoked will also be attributes of probability distributions.

The fact that statistical mechanical analogues of temperature and entropy are introduced that, unlike their thermodynamic counterparts, are attributes of probability distributions rather than of the physical states of systems, has given rise to some confusion. It might appear that we have changed the subject, and are no longer talking about individual physical systems but about ensembles of them, or that we are treating the probability distributions as representing the physical states of systems.[5] The reader will, I hope, be relieved to be assured that none of this is true. The subject matter is still individual physical systems and their properties. The shift that has been made is that, instead of making unqualified statements about the physical states of individual systems, we are making probabilistically qualified statements about them.

Nonetheless, the fact that statistical mechanical analogues of temperature and entropy are attributes, not of mechanical states of physical systems, but of *probability distributions* over mechanical states, is one that might seem surprising. One might antecedently have expected to find *mechanical* analogues of temperature and entropy, that is, quantities defined in terms of the physical properties of a system. An objection to this move that is often made is that, unlike the attributes of probability distributions that are the statistical mechanical analogues of temperature and entropy, thermodynamic temperature and entropy are *measurable* properties of a system. And this is true — in the quasi-deterministic regime! Outside of this regime, repetitions of a procedure we think of as a measurement of temperature (or entropy) will not yield a consistent value, but, rather, a distribution of values, and it makes no sense to talk of *the* value yielded by the procedure.

[5] This tendency is exacerbated somewhat by the fact that sometimes terminology is employed that has the potential to mislead, in that one talks of saying that probability distributions *representing* physical systems.

A quantity that has been regarded as a statistical mechanical analogue of entropy is the *Boltzmann entropy*, a generalization of the quantity H used by Boltzmann in his studies of the relaxation of a gas to equilibrium. In the quasi-deterministic regime, the difference between the Boltzmann entropies of two equilibrium states will generally approximate the difference between their thermodynamic entropies. Outside the quasi-deterministic regime, Boltzmann entropy does not enter into any useful analogue of the second law of thermodynamics. This is, of course, not to say that it isn't suited to the purpose for which it was introduced — as a quantity useful for tracking relaxation to equilibrium of an isolated system.[6]

I have been talking about the *relation* between thermodynamics and statistical mechanics, without commitment to whether that relation is one of reduction. Once one gets clear about the relation between thermodynamics and statistical mechanics, and how analogues of the laws of thermodynamics are obtained from statistical mechanics, the answer to the question of whether the relation is one of "reduction" depends, of course, on what is to count as a reduction. I will not, in this note, go into the matter of whether the relation between thermodynamics fits any of the models of reduction extant in the literature. It is almost a truism that the relation of thermodynamics to statistical mechanics is a relation of reduction if anything is. *If* one accepts that conditional — and it is by no means obvious that one should — and if one finds that the relation between thermodynamics and statistical mechanics doesn't satisfy one's favoured model of reduction, there are, of course, two possible responses to this situation. One would be to modify one's model of reduction, to accommodate the relation between thermodynamics and statistical mechanics. The other would be to draw an anti-reductionist moral.

[6] There is a tendency to conflate the second law of thermodynamics with the statement that systems, left to themselves, tend to relax to a state of equilibrium, a statement that Brown and Uffink (2001) have called the *minus first law*. This conflation should be resisted.

The tendency to conflate the two is encouraged by a tendency to treat the second law as a statement about the behaviour of isolated systems, to the effect that the entropy of an isolated system cannot decrease. This is a *consequence* of the second law, but it cannot be a statement of it, as the second law is a presupposition of the definition of thermodynamic entropy.

2 The laws of thermodynamics

The laws of thermodynamics with which we will be concerned are three: the zeroth law, the first law, and the second law. There is also a third law of thermodynamics, but it will not concern us today.

There is also another proposition that has been called a law of thermodynamics. This is what Brown and Uffink (2001) have called the *minus first law*, or *equilibrium principle*. It states that an isolated system in a confined space will eventually relax to an equilibrium state that is a function only of its internal energy, the external constraints (such as confinement to a container) imposed on it, and any external fields applied to it (such as gravitational or magnetic fields). Explanation of the tendency of systems to equilibrate has, understandably, been a major focus of the philosophical literature on the relations between statistical mechanics and thermodynamics, and our understanding of those relations is not complete until we have an explanation of that tendency. However, there are principled reasons for not counting the equilibrium principle as a law of thermodynamics.[7] The equilibrium principle, unlike the zeroth, first, and second laws, can be formulated without the distinction that is at the core of thermodynamics, between energy transfer as work and as heat. Thermodynamics presumes that systems do tend to equilibrate when left to themselves, but this can profitably be regarded as a *presupposition* of thermodynamics, rather than part of its subject matter. Statistical mechanics should, of course, take as part of its task to explain why we might be justified in supposing this, but this can be regarded as a task distinct from that of obtaining analogues of the laws of thermodynamics.[8]

2.0 The zeroth law

The zeroth law has to do with relations between systems that are in thermal equilibrium and have been placed in thermal contact with each other. It thus presupposes that we understand what it means for a system to be in thermal equilibrium, and what it means for two systems to be in thermal contact — that is, to be in a situation that permits heat flow between the systems.

[7] And, as a matter of historical fact, though something of the sort was long explicitly recognized as an important principle, it was not counted by anyone as a law of thermodynamics until the 1960s, more than a century after Kelvin initiated talk of laws, or fundamental principles, of thermodynamics.

[8] See Myrvold (2020a) for further discussion of this point.

When two systems are placed in thermal contact, there may be heat flow from one to the other, or there may be none. The matter of whether or not there will be heat flow, and, if there is, in which direction, is a relation between the thermodynamic states of the systems. We're interested in the case in which there is no heat flow when the bodies are placed in thermal contact. This relation between themodynamic states is obviously a symmetric one. The zeroth law says that it is transitive.

> **The zeroth law of thermodynamics, version I.** If a and b are two thermodynamic states such that, if systems in these states are brought into thermal contact, there is no net heat flow between them, and if b and c are two thermodynamic states for which this is also true, then the same holds for a and c.

If the zeroth law holds, then we can partition thermodynamic states into equivalence classes, which we will regard as classes of states that all have the same temperature. Therefore, another way of putting the zeroth law, which we will take as our preferred expression of it, is the following.

> **The zeroth law of thermodynamics, version II.** There is a function T of thermodynamic states, such that, if systems in state a and b are brought into thermal contact, there will be no net heat flow between them if and only if $T(a) = T(b)$.

2.1 The first law

Thermodynamics is based on the realization that there is a mechanical equivalent of heat. That is, when work is expended to create heat, the same amount of work is required to produce a given quantity of heat, regardless of how this is done, and, when heat transfer is exploited to produce mechanical work, the amount of heat that produces a given quantity of work is always the same. It is this that permits us to speak of the total internal energy of a system, which can be changed either by flow of heat in or out of the system, or by mechanical means, by doing work on the system or having it do work on some external system.

> **The first law of thermodynamics.** There is a function U of thermodynamic states, the internal energy of the system, such that, in any process, the change in U is equal to the sum of the

net heat that passes in or out of the system, and the net work done on or by the system.

2.2 The second law

The second law has been stated in a variety of forms. One is the *Clausius version*, which states that spontaneous flow of heat is from warmer to colder bodies, and that, furthermore, there is no process that has no net effect other than moving heat from a cooler to a warmer body. It follows from this that there is an upper bound on the efficiency of any heat engine operating in a cycle between two reservoirs (that is, work developed per unit of heat obtained from the warmer reservoir). This permits the definition of an absolute temperature scale. If η_{12} is the upper bound on efficiency of an engine operating between two reservoirs, we can define the ratio of the absolute temperature of the warmer, T_1, to that of the cooler, T_2 by,

$$\frac{T_2}{T_1} = 1 - \eta_{12}. \tag{1}$$

This permits us to formulate what we will take to be our preferred statement of the second law of thermodynamics.

> **The second law of thermodynamics** It is possible to choose a temperature function (that is, a state function that takes equal values on states of equal temperature) in such a way that there is a function S of thermodynamic states of a system with the property that, in any process that takes a system from a state a to state b, exchanging heats Q_i with heat reservoirs at temperatures T_i,
>
> $$\sum_i \frac{Q_i}{T_i} \leq S(b) - S(a).$$

If the states a and b can be connected reversibly — that is, if there is a process that takes a to b, and another process that takes b to a with the signs of the heat exchanges reversed — then the function S of which the second law speaks is fixed up to an additive constant, and we can define differences in thermodynamic entropy S_θ, by

$$S_\theta(b) - S_\theta(a) = \int_a^b \frac{đQ}{T}, \tag{2}$$

where the integral can be taken over any reversible process.

This also works if reversibility is unachievable but can be approached arbitrarily closely; that is, if, for any $\varepsilon > 0$, there are processes that take a to b and then b to a such that $\int đQ/T$, taken over the full process, is less than ε. We can then define $S_\theta(b) - S_\theta(a)$ as the least upper bound of the set of values of the integral of $đQ/T$, taken over all processes.

It is the second law, together with the condition that any two thermodynamic states can be connected reversibly or arbitrarily close to reversibly, that permits the assignment of a unique entropy difference to any pair of thermodynamic states. If the second law fails, there is no state-function that bounds heat exchanges in the way demanded in our statement of the second law. If reversibility fails, and if there are limits on how closely reversibility can be approximated, there will be a multitude of functions that fit the bill. This will be important, later, for our discussion of statistical mechanical analogues of thermodynamic entropy.

3 Statistical mechanical analogues of the laws of thermodynamics

What has to change, once we acknowledge that the variables that we use to characterize the thermal state of a system fall radically short of the full set of variables potentially relevant to prediction of its behaviour?

First, and foremost, we must acknowledge that what the zeroth and second laws say *must* happen, may be liable to exceptions, albeit ones that on the macroscopic scale are almost sure to be insignificant. In this vein, we find Maxwell, in a letter to John Strutt, Baron Rayleigh, of Dec. 6, 1870, drawing the

> *Moral.* The 2nd law of thermodynamics has the same degree of truth as the statement that if you throw a tumblerful of water into the sea, you cannot get the same tumblerful of water out again. (Garber et al. 1995, p. 205; Harman 1995, p. 583).

Gibbs drew a similar conclusion, several years later.

> when such gases have been mixed, there is no more impossibility of the separation of the two kinds of molecules in virtue of their ordinary motions in the gaseous mass without any external

> influence, than there is of the separation of a homogeneous gas into the same two parts into which it sas once been divided, after these have once been mixed. In other words, the impossibility of an uncompensated decrease of entropy seems to be reduced to improbability (Gibbs 1875, p. 229, in Gibbs 1906, p. 167).

Maxwell elaborated on this point, in his review of Tait's *Sketch of Thermodynamics*.

> If we restrict our attention to any one molecule of the system, we shall find its motion changing at every encounter in a most irregular manner.
>
> If we go on to consider a finite number of molecules, even if the system to which they belong contains an infinite number, the average properties of this group, though subject to smaller variations than those of a single molecule, are still every now and then deviating very considerably from the theoretical mean of the whole system, because the molecules which form the group do not submit their procedure as individuals to the laws which prescribe the behaviour of the average or mean molecule.
>
> Hence the second law of thermodynamics is continually being violated, and that to a considerable extent, in any sufficiently small group of molecules belonging to a real body. As the number of molecules in the group is increased, the deviations from the mean of the whole become smaller and less frequent; and when the number is increased till the group includes a sensible portion of the body, the probability of a measurable variation from the mean occurring in a finite number of years becomes so small that it may be regarded as practically an impossibility.
>
> This calculation belongs of course to molecular theory and not to pure thermodynamics, but it shows that we have reason for believing the truth of the second law to be of the nature of a strong probability, which, though it falls short of certainty by less than any assignable quantity, is not an absolute certainty (Maxwell 1878, p. 280; Niven 1890, pp. 670–71).

Boltzmann also acknowledged the point, in his response to Loschmidt, who had drawn his attention to the fact that, because of the invariance of the laws of mechanics under the operation of velocity reversal, no temporally asymmetric conclusion can be drawn from mechanical considerations alone.

> Loschmidt's theorem seems to me to be of the greatest importance, since it shows how intimately connected are the second law and probability theory, whereas the first law is independent of it. In all cases where $\int dQ/T$ can be negative, there is also an individual very improbable initial condition for which it may be positive; and the proof that it is almost always positive can only be carried out by means of probability theory (Boltzmann 1966, p. 189, from Boltzmann 1877, in Boltzmann 1909, p. 121)

If we are to take these considerations seriously, we must incorporate probabilistic considerations into our reasoning, and seek probabilistic counterparts to the zeroth and second laws, according to which what the original, thermodynamic versions of these laws declare to be impossible, is at most improbable.

Another change is a reconceptualization of equilibrium. Thermodynamic equilibrium is, to the unaided eye, a state of tranquility. At the molecular level it is, of course, seething with activity, and at a mesoscopic scale some measurable variables fail to settle into a steady state. The best-known example of this is, of course, Brownian motion, invisible to unaided eyes, but visible with a microscope of modest power.

One could choose to say that a system of that sort, for which observable parameters refuse to settle down to steady values, fails to equilibrate. But we should not ignore the fact that a system like that, if disturbed, settles down to a condition in which the observable parameters exhibit a regular pattern of fluctuations, to which we can attach a well-defined probability distribution. For this reason, it makes sense to talk about an equilibrium *distribution*, which is approached as a system equilibrates. Only in the quasi-deterministic regime will the equilibrium distribution be such that the observable parameters take precise values (within the limits of observation), from which they will with extremely high probability not depart appreciably on the time scales of observation, and we can talk about an equilibrium *state*.[9]

[9] Thus, speaking of equilibrium distributions, as is done in "Gibbsian" statistical mechanics, and speaking of equilibrium states, as one does in "Boltzmannian" statistical mechanics, does not reflect rival, incompatible conceptions of equilibrium; the latter is a special case of the former.

3.0 Statistical mechanical analogue of the zeroth law

Consider first, the zeroth law. When two systems at the same temperature are placed in thermal contact, we expect that, on a sufficiently fine-grained level, the energies of the two systems will not be strictly constant, but will fluctuate. The zeroth law is, therefore, not quite right. But, on a macroscopic scale, it will be *close* to being right. For two macroscopic bodies of the same temperature in thermal contact with we expect the energy fluctuations of each body to be minuscule compared to the total energy of the body, and, furthermore, that the energy flow will go either way, with no net tendency in either direction, and that the average energy exchange, over any sufficiently long time period, will almost certainly be close to zero.

This suggests, of course, that the zeroth law be replaced with a probabilistic version, and that, rather than attempting to make unqualified statements about what will happen when two systems are brought into thermal contact, we make probabilistically qualified statements. The version that we will adopt will, in fact, refer to *expectation values* of energy exchanges.

The zeroth law has to do with systems that have relaxed to thermal equilibrium. Given macroscopically available information, we will not be able to make unqualified, categorical assertions about physical parameters such as its energy, but we may be in a position to assign more or less definite probabilities to ranges of values of those parameters. A family of probability distributions that play a key role are the *canonical distributions*. These are indexed by a parameter β. A canonical distribution with index β is a probability distribution on the phase space of system that has a density function with respect to Liouville measure,

$$\tau_\beta(x) = {Z_\beta}^{-1}\, e^{-\beta H(x)}. \tag{3}$$

A widely accepted postulate of statistical mechanics is the following.

> **The Canonical Postulate.** For a system that has relaxed to thermal equilibrium, a canonical distribution, uncorrelated with the system's environment, is appropriate for the purposes of making probabilistic assertions about the values of macrovariables or the system's responses to manipulations of macrovariables.

It will usually not be expressed quite like this, but if one pays attention to the uses made of canonical distributions, one will see that this, or something much like it, is the assumption being made.

The Canonical Postulate is the key to obtaining statistical mechanical analogues of the laws of thermodynamics. Once we have it, the statistical mechanical analogues of thermodynamical laws are relatively simple theorems. The postulate is not, of course, the sort of thing that should simply be accepted without good reason. Support for the Canonical Postulate comes from a combination of empirical and theoretical considerations, some of which involve investigation of the process of equilibration. In this note we will not go into the reasons for accepting the Canonical Postulate, but will confine ourselves to exploring its consequences.[10]

We will actually be needing only a weaker version; the only macrovariables whose distributions we will need is the exchange of energy of the system with bodies placed in thermal contact with it.

Thermal contact between two systems will be represented by a coupling of the systems such that the contribution to the total internal energy of the pair of systems due to the coupling is small compared to the internal energy of each system, so that the total internal energy of the pair of systems is approximately equal to the sum of the internal energies of the two systems. We also assume that a pair of systems can be thermally coupled and decoupled without doing any work on the joint system, that is, without changing the total energy of the pair of systems.

Canonical distributions have the following useful property, which is part of the reason why they are thought to be appropriate for thermal states. When two initially uncoupled systems with which are associated canonical distributions with parameters β_a and β_b are allowed to interact *in any way whatsoever*, with the proviso that in the process of coupling and decoupling them no work is done on the joint system and the total energy is unchanged, the direction of the expectation value of energy flow between the two systems is determined *solely* by which of the parameters is larger, and is independent of the details of the coupling It can also be shown that, under the same conditions (which involve no work done on the joint system during the process of coupling and decoupling), if the initial parameter values are the same, the expectation value of energy exchange is zero.

Proposition 1. *Suppose that, at time t_0 two systems A and B have independent canonical distributions τ_{β_a} and τ_{β_b}. Between time t_0 and t_1 they interact with each other in such a way that the total energy of the pair of systems is*

[10] See Maroney (2007) for an illuminating discussion of the rationale for the Canonical Postulate.

conserved. At t_1 they are again no longer interacting. Let $\langle \Delta H_A \rangle$ be the expectation value of the energy change of system A (which is, of course, the negative of $\langle \Delta H_B \rangle$, the expectation value of the energy change of system B). Then,

- *If $\beta_a = \beta_b$, $\langle \Delta H_A \rangle = 0$,*
- *if $\beta_a > \beta_b$, $\langle \Delta H_A \rangle \geq 0$, and*
- *if $\beta_a < \beta_b$, $\langle \Delta H_A \rangle \leq 0$.*

It will sometimes be convenient to work with a parameter T, related to β via

$$\beta = \frac{1}{kT}, \tag{4}$$

where k is a constant, called *Boltzmann's constant*, which, as we will see in section 4, is related to the constant R that appears in the ideal gas law via Avogadro's number. Then we can restate Proposition 1 as,

Proposition 2. *Under the same conditions as in Proposition 1,*

- *If $T_a = T_b$, $\langle \Delta H_A \rangle = 0$,*
- *if $T_a < T_b$, $\langle \Delta H_A \rangle \geq 0$, and*
- *if $T_a > T_b$, $\langle \Delta H_= A \rangle \leq 0$.*

The Canonical Postulate and Proposition 2 yield our statistical mechanical analogue of the zeroth law of thermodynamics.

> **The zeroth law of thermodynamics, statistical mechanical analogue.** The class of probability distributions appropriate to making probabilistic statements about macroscopic behaviour of systems that have relaxed to thermal equilibrium is indexed by a parameter T, such that, if two such systems, with which are associated probability distributions for index-values T_a and T_b, are brought into thermal contact with each other, the *expectation value* of energy flow between the systems is zero if and only if $T_a = T_b$.

Compare this with the original, thermodynamic version of the second law. The statistical mechanical version is obtained from the thermodynamic version via the substitution of equilibrium probability distributions for thermodynamic states, and expectation values of energy flow for actual energy flows. In the quasi-deterministic regime, in which expectation values of energy exchanges may be taken for the actual values, we recover the original version. But the statistical mechanical version holds even when energy fluctuations are large, so that the actual value of energy flow need not be even *probably* close to its expectation value.

3.1 Statistical mechanics and the first law

Once we have settled on how to distinguish between energy transfer as work and energy transfer as heat, the first law of thermodynamics is just a statement of the conservation of energy. The concept of energy, unlike the concepts of temperature and entropy, is a concept that belongs to the underlying mechanics, and its conservation a consequence of the microphysical laws.

Though energy is a concept of the underlying mechanics, the distinction between work and heat is not. This raises the question of how to distinguish between work and heat, in the context of statistical mechanics.

The standard answer goes as follows. Some of the variables that define the state of a system are treated as exogenous, manipulable variables.[11] Energy changes to the system via manipulation of these variables are to be counted as work; energy changes via interactions with a heat reservoir, as heat.[12]

To get a feeling for how this works, consider a system in contact with a heat reservoir. The total Hamiltonian of the joint system, consisting of the system under consideration and the heat reservoir, is a function of the phase-space point of the combined system, and the exogenous parameters.

Suppose that at some time t_0, the system is in thermal equilibrium with a heat reservoir with parameter β_a, and is subject to external parameters $\boldsymbol{\lambda} = \{\lambda_1, \ldots, \lambda_n\}$. At t_1 it is subject to parameters $\{\lambda_1 + d\lambda_1, \ldots, \lambda_n + d\lambda_n\}$ that differ, if at all, by small amounts from their original values, and is in thermal equilibrium with a heat bath with parameter $\beta = \beta + d\beta$. In accordance with the Canonical Postulate, we use canonical distributions with the respective

[11] See Myrvold (2020b) for discussion of what this entails.

[12] The astute reader will have noticed that this need not be an exhaustive partition. This is correct. For the purposes of this note, however, we will confine ourselves to situations in which there are no other sorts of interaction between the system and its environment.

parameters to calculate probabilities for properties of the system at each of these times. It is not difficult to show that the expectation value of the change in total energy of the system, between times t_0 and t_1, is,

$$\langle dU \rangle = -kTd\,\langle \log \rho \rangle + \sum_{i=1}^{n} \left\langle \frac{dH}{d\lambda_i} \right\rangle d\lambda_i, \tag{5}$$

where H is the Hamiltonian (total energy) function of the system, and ρ is a probability density with respect to Liouville measure. The second term on the right-hand-side of (5) is the expectation value of work done in changing the external parameters; the first term is, therefore, the expectation value of heat exchanged with the reservoir. Thus, we have up to first order in parameter differences,

$$\langle đQ \rangle = -kT\, d\langle \log \rho \rangle. \tag{6}$$

3.2 Statistical mechanical analogue of the second law

The original version of the second law places limits on the efficiency with which a heat engine operating in a cycle between two reservoirs, at temperatures T_1 and T_2, can operate. If Q_1 is the heat obtained from the hotter reservoir, at temperature T_1, and W is the work obtained in a cycle, it follows from the second law that

$$W \leq \left(1 - \frac{T_2}{T_1}\right) Q_1. \tag{7}$$

This is the Carnot bound on the efficiency of heat engines.

Because of fluctuations in the energy transferred, (7) might not hold, on a given run of the engine. We *might* get more work than the Carnot bound allows. Over the years various schemes have been proposed that are meant to exploit molecular-scale fluctuations to create a perpetual motion machine of the second kind, which could operate in a cycle between two heat reservoirs to consistently and reliably exceed the Carnot bound on efficiency. These all fall afoul of the unpredictability of fluctuations. Though, on one occasion, we might get more work in a cycle than allowed by the Carnot bound, we also might get less. The question is whether we can expect to get more, on average, work than permitted by the Carnot bound.

One thing we might expect from a statistical mechanical version of the second law would be a probabilistic statement to the effect that no engine can consistently and reliably violate the Carnot bound. Szilard compared with this with no-go theorems for gambling schemes that seek to consistently and reliably beat the odds at a casino.

> Consider somebody playing a thermodynamical gamble with the help of cyclic processes and with the intention of decreasing the entropy of the heat reservoirs. Nature will deal with him like a well established casino, in which it is possible to make an occasional win but for which no system exists ensuring the gambler a profit (Szilard 1972, p. 73, from Szilard 1925, p. 757).

This suggests that we might be able to obtain a probabilistic version of the second law that yields a Carnot bound on *expectation value* of work obtained. That is, something of the form,

$$\langle W \rangle \leq \left(1 - \frac{T_2}{T_1}\right) \langle Q_1 \rangle. \tag{8}$$

And, indeed, if one accepts the Canonical Postulate, which tells us what probability distributions are appropriate for heat reservoirs, we get just that.

> **Statistical second law of thermodynamics**
> *a. Classical.* There is a functional S of probability distributions, such that: If a system A that, at time t_0, has associated with it a probability distribution $\rho_A(t_0)$, interacts in the interval between t_0 and t_1 with heat reservoirs B_i that, at t_0, have canonical distributions τ_{β_i}, uncorrelated with A, then the expectation value of heat exchanges Q_i between A and the reservoirs are bounded by,
>
> $$\sum_i \frac{\langle Q_i \rangle}{T_i} \leq S[\rho_A(t_1)] - S[\rho_A(t_0)].$$
>
> *b. Quantum.* The same, with quantum states replacing probability distributions.

The reader should take a moment to verify that this does, indeed, entail the probabilistic version of the Carnot bound, inequality (8).

If $\rho_A(t_1)$ and $\rho_A(t_0)$ can be connected reversibly — that is, if there is some process that takes $\rho_A(t_1)$ back to $\rho_A(t_0)$, in the course of which there are heat exchanges with the heat reservoirs whose expectation values are those of the process that takes $\rho_A(t_0)$ to $\rho_A(t_1)$, with signs reversed — then the functional S of which the statistical second law speaks is uniquely determined, up to an additive constant. If there are bounds on how closely reversibility can be approached, there will be a plethora of such functionals, whose differences are not constant.

The statistical second law is a *theorem* of statistical mechanics. It depends on associating canonical distributions with the heat reservoirs, but there is no assumption about the initial probability distribution for A, $\rho_A(t_0)$, other than that the state of A is uncorrelated with that of the heat reservoirs. The statistical second law follows from what I have elsewhere called the *Fundamental Theorem of Statistical Thermodynamics* (see Myrvold 2020b, 2021b).

Proposition 3. *a. Classical. Consider a system A that, at time t_0, has associated with it a probability distribution $\rho_A(t_0)$ that has a density f_ρ with respect to Liouville measure. In the interval between t_0 and t_1 it interacts with heat reservoirs B_i that, at t_0, have canonical distributions τ_{β_i}, uncorrelated with A. Then the expectation value of heat exchanges Q_i between A and the reservoirs are bounded by,*

$$\sum_i \frac{\langle Q_i \rangle}{T_i} \leq S[\rho_A(t_1)] - S[\rho_A(t_0)],$$

where S_G is the "Gibbs entropy,"

$$S_G[\rho] = -k \langle \log f_\rho \rangle_\rho.$$

b. Quantum. The same, with a quantum state represented by a density operator $\hat{\rho}$ replacing the probability distribution, and with the functional being the von Neumann entropy*,*

$$S_{vN}[\hat{\rho}] = -k \langle \log \hat{\rho} \rangle_\rho.$$

Proof is found in the Appendix.

3.3 Thermodynamically equivalent distributions, and coarse-grained entropies

Consider the familiar example of free expansion of a gas composed of macroscopically many molecules. At time t_0, the gas is confined to one side of a container by a partition. The partition is removed, and the gas is allowed to diffuse throughout the container. Let t_1 be some time after the removal of the partition, long enough that the gas has had ample time to equilibrate. For illustrative purposes, we consider a wholly fictitious situation in which the gas undergoes isolated evolution during this time, free from any external influences.

We expect that the gas, at t_1, will be fairly evenly distributed throughout the box. That is, if we partition the box into subvolumes, which may be small compared to the total volume of the box but must be large enough to contain a great many molecules, we will expect the amount of gas in each subvolume to be approximately proportional to its volume. Because the gas undergoes isolated evolution, its energy at the end of the process is the same as it is at the beginning. If the gas is sufficiently rarified that its behaviour approximates that of an ideal gas, to which Joule's law applies, the temperature is a function only of its internal energy, and is the same at the beginning and end of the process. For simplicity, we will assume that this is the case in the discussion that follows.

Let us ask: is the entropy at the end of the process higher, lower, or the same as at the beginning? To answer this question, thermodynamics requires us to answer the question of whether the process of free expansion, during which the gas exchanges no energy with the external world, is to be counted as a reversible process or not. If there were a device that could instantaneously reverse the velocities of all the molecules in the gas, with no other effect on the gas, and without expenditure of work, then the process would be reversible, and we would have to count the initial and final states as having the same thermodynamic entropy.[13]

There is no such device, of course, and free expansion of a gas is not even close to being reversible. A process that *is* close to reversible is a process of gradual expansion of the gas against a piston that at each moment exerts a force on the gas that is approximately equal to the force exerted by the pressure of the gas on the piston, during the course of which the gas is kept at constant temperature by thermal contact with a heat bath. As the gas does work on the piston, imparting energy to it, it absorbs compensating quantities of heat from the heat bath. There is in the course of this process a positive net influx of heat into the gas, and, therefore, a positive change of

[13] If this remark is surprising to you, then review the definition of thermodynamic entropy, as this is an immediate consequence of that definition.

entropy, and the final entropy is higher than the initial entropy.

This leads to the conclusion that the final state of a gas that undergoes a gradual expansion of this sort is a state of higher entropy than the initial state? But what implications does this have for the final state of a gas undergoing free expansion, while isolated from its environment?

The answer is, of course, that we count the final states of these two very different processes as instances of the *same thermodynamic state*, and *ipso facto* assign the same entropy to the end state, regardless of which process leads up to it. The rationale for this lies in the expectation that the process of equilibration that the gas undergoes during free expansion effectively effaces all traces of its recent past, and that, as far as measurements of macroscopically accessible quantities, and the responses of the gas to feasible manipulations, are concerned, there is no difference between a gas that has recently relaxed, via an irreversible process, from a far-from equilibrium state, and a gas with the same volume and temperature that has been in thermal contact with a heat bath for a protracted period of time. It is as if the gas has drunk from the river Lethe and has forgotten its past (or rather, has suppressed the memory, because, in the fictitious situation we are imagining, of completely isolated evolution, the traces of the past remain, but are embedded in details of the microstate that are macroscopically inaccessible).

This is a common assumption of thermodynamics, more often implicitly invoked than made explicit. But let's make the assumption explicit.

> **The Lethean postulate.** As far as results of measurements of macroscopically accessible quantities are concerned, and responses to feasible manipulations, a system that has recently relaxed to equilibrium in isolation from its environment is indistinguishable from a system with the same values of external constraints (such as volume), and same temperature, that has been in protracted thermal contact with a heat bath at that temperature.

In what follows, we assume that the Lethean postulate is correct, and explore its consequences for the question of statistical mechanical analogues of thermodynamic entropy.

Let τ_{β,V_0} be a canonical distribution for some temperature $T = 1/(k\beta)$ over the portion of the container available to the gas at time t_0, and let $\rho(t_1)$ be the result of evolving this distribution, via the dynamics of the gas, from

t_0 to t_1. Let τ_{β,V_1} be a canonical distribution, with the same temperature, over the volume of the container available to the gas at t_1.

These two probability distributions $\rho(t_1)$ and τ_{β,V_1} will be similar in some respects, and very different in others. Because the evolution from t_0 to t_1 conserves energy, they will agree on the distribution of energy; that is, they will agree not only on the mean value of energy, but also on the extent of the spread in energy. They will also agree closely on probabilities regarding the amount of gas to be found in any subregion of the box; both will accord high probability to the gas being spread fairly evenly throughout the box, with the amount of gas in any not-too-small subvolume proportional to its volume. In fact, they will agree very nearly on the probability distribution of any macroscopically measurable parameter.

They will disagree on other matters. $\rho(t_1)$ has support on the set of states that can be reached, in the course of isolated evolution, from the set of states accessible at t_0. This is a set that is accorded probability 1 by $\rho(t_1)$ and very small probability, of the order $(V_0/V_1)^N$, where N is the number of molecules in the gas, by τ_{β,V_1}. This set is very finely distributed throughout the region of phase space available at t_1. Thus, $\rho(t_1)$, unlike τ_{β,V_1}, sharply distinguishes between some sets of phase space that differ in ways that are invisible to macroscopic scrutiny.

Suppose that, after time t_1, we want to restore the system that has been allowed to relax to equilibrium while isolated from its surroundings to a state in which it is at its original temperature, is confined to its original volume and is in contact with a heat bath at that temperature. We have available to us various heat reservoirs, at temperatures T_i, and we want to invoke the statistical-mechanical analogue of the second law to place bounds on expectation values $\langle Q_i \rangle$ of heat exchange with these reservoirs in the course of the transition. The statistical second law tells that, no matter how we achieve the transition,

$$\sum_i \frac{\langle Q_i \rangle}{T_i} \leq S[\tau_{\beta,V_0}] - S[\rho(t_1)]. \tag{9}$$

There's a wrinkle, however. Because $\rho(t_1)$ is obtained from τ_{β,V_0} via isolated evolution, the two distributions have the same Gibbs entropy, and so (9) tells us

$$\sum_i \frac{\langle Q_i \rangle}{T_i} \leq 0. \tag{10}$$

This is correct, but is consistent with the quantity on the left-hand side being zero, or arbitrarily close to zero. We know, however, that, in order to compress the gas to its original volume without changing from temperature, some heat must be expelled from the gas.

If, as we believe, the Lethean Postulate is correct, we can do better, and achieve a tighter bound. If the Lethean postulate is correct, the fine-grained details of $\rho(t_1)$ are irrelevant to expectation values of heat exchanges over the course of any process due to manipulations of macroscopic variables, and these expectation values are the same whether calculated with respect to $\rho(t_1)$ or τ_{β,V_1}, and so we can conclude that, over the course of any process involving macroscopic manipulations that take the system back to its original state,

$$\sum_i \frac{\langle Q_i \rangle}{T_i} \leq S[\tau_{\beta,V_0}] - S[\tau_{\beta,V_1}] = -kN \log(V_1/V_0), \tag{11}$$

which is a much better bound!

This is an illustration of the remark made earlier, about the uniqueness (up to an additive constant) or lack thereof of entropy functions. If any two thermodynamic states can be connected reversibly, the state-function whose existence is asserted in the second law is uniquely determined, up to an arbitrary constant. If there are states that cannot be connected reversibly, then there will be more than one function that plays the role of placing bounds on heat exchanges. This is one such case. There is no process of macroscopic manipulation that takes the probability distribution $\rho(t_1)$ back to $\rho(t_0)$ without expenditure of work. Such a process would have to exploit the fine-grained differences between $\rho(t_1)$ and τ_{β,V_1}, and would have to be something akin to a magical reversal of all the velocities of the gas molecules. Thus, the process of expansion does not count as reversible, and we are in a situation in which the second law does not uniquely determine a function S that bounds the expectation values of heat exchanges. When such a variety of bounds exist, the bound of most interest is the strongest bound, which entails all the others.[14]

The probability distribution τ_{β,V_1} is a "smoothing" of $\rho(t_1)$, in the sense that they agree on probability distributions of macroscopically ascertainable variables, but $\rho(t_1)$ makes discriminations between macroscopically indistinguishable microstates that τ_{β,V_1} doesn't, as τ_{β,V_1} has a density function that

[14] For a systematic discussion of how to do thermodynamics and statistical thermodynamics without reversibility, see Myrvold (2020b).

is a function only of macrovariables. Let us generalize this smoothing operation. Given a set $\{F_1, \ldots, F_n\}$ of macrovariables, let us say that a probability distribution ρ is a *smoothing* (or *coarse-graining*) of a distribution ρ, with respect to these macrovariables, if $\bar{\rho}$ agrees with ρ on distributions for values of the macrovariables $\{F_1, \ldots, F_n\}$, and $\bar{\rho}$ has a density function that is a function only of the variables $\{F_1, \ldots, F_n\}$.

Let us say that two probability distributions are *thermodynamically equivalent* if they agree on probability distributions of all macroscopically accessible variables and also agree on the expectation values of heat exchanges with any heat baths that result from macroscopic manipulations. Then one way to express the Lethean postulate is that any probability distribution is thermodynamically equivalent to its smoothing $\bar{\rho}$ with respect to all relevant macrovariables.

Armed with this, we get the following strengthening of our Fundamental Theorem.

Proposition 4. *a. Classical. Suppose a system A, has associated with it at time t_0 a probability distribution $\rho_A(t_0)$ that has a density f_ρ with respect to Liouville measure. In the interval between t_0 and t_1 it interacts with heat reservoirs B_i that, at t_0, have canonical distributions τ_{β_i}, uncorrelated with A. If $\rho_A(t_0)$ is thermodynamically equivalent to $\bar{\rho}_A(t_0)$, then the expectation value of heat exchanges Q_i between A and the reservoirs are bounded by,*

$$\sum_i \frac{\langle Q_i \rangle}{T_i} \leq S_G[\rho_A(t_1)] - S_G[\bar{\rho}_A(t_0)],$$

where, again, S_G is the "Gibbs entropy."
b. Quantum. Same, but with quantum states and von Neumann entropy.

A common refrain of polemics directed against "Gibbsian" methods is that, for a system that equilibrates in isolation from its environment, the fine-grained Gibbs entropy fails to represent the difference in thermodynamic entropy between initial and final states. This is correct; it does not. This would count as an objection to anyone who has claimed that it does, but one may search the literature in vain for a claim of that sort. It is also commonly said that invocation of coarse-grained Gibbs entropies is unmotivated. I hope that the above discussion has made it clear that this is simply and plainly false. We set out to look for an analogue, in statistical mechanical terms, of the second law of thermodynamics, that places bounds on expectation values of heat exchanges in terms of some functional of probability distributions. If two probabilistic "states" can be connected reversibly, the requirements imposed by the theorem determine, up to an additive constant, what that functional is. In the absence of reversibility, more than one functional takes this role, but the strongest bound is the one of most interest, and entails all of the others. If the Lethean postulate holds, we will, whenever the coarse-grained and fine-grained entropies of the initial probability distribution differ, obtain a stronger bound if we invoke coarse-grained entropy.

4 Measured values of temperature and entropy

A question that has undoubtedly formed in the reader's mind is: what is the relation between the attributes of probability distributions, introduced in the previous section, and macroscopically measurable values of temperature and entropy? The answer has to do with the quasi-deterministic regime.

Let us consider the case of temperature, first. The simplest thermometer to analyze is a constant pressure ideal gas thermometer. Take a monatomic ideal gas consisting of N molecules, enclosed in a container whose volume can change (for simplicity, one might imagine a cylinder with a movable piston), whose expansion is resisted by a constant pressure p on the outside of the container. The volume is taken as an indicator of temperature, via the ideal gas law,

$$pV = nRT, \tag{12}$$

where R is the ideal gas constant, and n is the number of moles of gas contained in the thermometer, that is, the number of molecules, N, divided by Avogadro's number, N_A. So, if we put the thermometer in thermal contact with some body, and allow the joint system to equilibrate, the temperature measured by our gas thermometer, as indicated by its volume V, is

$$T_{meas} = \frac{pV}{nR} = \frac{pV}{N(R/N_A)}. \tag{13}$$

Suppose the thermometer is placed in thermal contact with some object. The joint system is allowed to equilibrate. In accordance with the Canonical Postulate, it will settle into a condition in which the two systems are exchanging quantities of energy and the energy of each separate system is a fluctuating quantity with probabilities given by a canonical distribution for the joint system, with some parameter β. This gives us a probability distribution for the volume, V, and, hence, a probability distribution for the measured temperature, T_{meas}.

For our gas thermometer, a canonical distribution with parameter β yields a probability distribution for the volume V with expectation value $\langle V \rangle$ that satisfies,

$$p \langle V \rangle = \frac{N+1}{\beta}, \tag{14}$$

and a standard deviation, $\sigma(V)$, given by,

$$\sigma(V) = \frac{\langle V \rangle}{\sqrt{N+1}}. \tag{15}$$

Therefore, for a macroscopic gas thermometer containing a number of molecules on the order of Avogadro's number, the spread in the probability distribution for V is of negligible size, compared to typical values of V. We are thus in the quasi-deterministic regime, and one may take the expectation value as the value that will almost certainly be obtained. The same holds, of course, for the measured temperature, T_{meas}; for a macroscopic thermometer, it has a probability distribution tightly focussed around its expectation value $\langle T_{meas} \rangle$. That expectation value is related to the parameter β by,

$$\langle T_{meas} \rangle == \frac{p\langle V \rangle}{N/(R/N_A)} = \left(1 + \frac{1}{N}\right) \frac{1}{(R/N_A)\beta}. \tag{16}$$

Take $k = R/N_A$. Recall that the temperature T was related to the parameter β by $\beta = 1/(kT)$. This gives us, for expectation value of the measured temperature,

$$\langle T_{meas} \rangle = \left(1 + \frac{1}{N}\right) T. \tag{17}$$

The standard deviation of the measured temperature is

$$\sigma(T_{meas}) = \frac{\langle T_{meas} \rangle}{\sqrt{N+1}}. \tag{18}$$

Thus, for large N, the temperature of an object, as measured by an ideal gas thermometer, will almost certainly closely approximate the parameter T invoked in the statistical version of the zeroth law.

What about entropy? The difference in thermodynamic entropy between two thermodynamic states, a and b, is calculated by considering some reversible process that links the state. Consider the statistical mechanical treatment of a system whose temperature and external parameters are slowly varied from $(\beta_a, \boldsymbol{\lambda}_a)$ to $(\beta_b, \boldsymbol{\lambda}_b)$, while the system is in thermal contact with a heat bath whose temperature is also slowly varied. If the time-scale of the change of parameters is slow compared to the equilibration time-scale of the system, we may take the system to be effectively in equilibrium at each stage of the process. According to the Canonical Postulate, we may use a family of canonical distributions with varying parameters to calculate probabilities for energy exchanges between the system and the heat bath. From (6),

$$\begin{aligned} \int_a^b \frac{\langle Q \rangle}{T} &= -k \left(\langle \log \rho_b \rangle_b - \langle \log \rho_a \rangle_a \right) \\ &= S_G[\tau_{\beta_b, \boldsymbol{\lambda}_b}] - S_G[\tau_{\beta_a, \boldsymbol{\lambda}_a}]. \end{aligned} \tag{19}$$

In the quasi-deterministic regime, actual energy exchanges will be close to their expectation values, so the measured entropy difference will, with high probability, be close to the difference in Gibbs entropies.

5 Boltzmann entropy and its relation to thermodynamic entropy

The quantity that has come to be called "Boltzmann entropy" is defined as follows. One chooses a number of functions on phase space $\{F_1, \ldots, F_n\}$, to be regarded as *macrovariables*. One then coarse-grains the ranges of these macrovariables, into intervals small enough that all values within the interval can be consider effectively the same, on a macroscopic scale. This gives a partitioning of the phase space of the system into *coarse-grained macrostates*. Typically one of the macrovariables is energy. Partitioning the range of possible energies divides the phase space of the system into "energy shells," consisting of a narrow band of energies, each of which is partitioned by the coarse-graining of the other macrovariables into macrostates. Let $M(x)$ be the element of the macrostate partition to which a phase-space point belongs, and let μ be Liouville measure on phase space. Then the Boltzmann entropy of a microstate x is defined as,

$$S_B(x) = k \log \mu(M(x)). \tag{20}$$

The connection between Boltzmann entropy and thermodynamic entropy is via Gibbs entropy. As we have seen in the previous section, in the quasi-deterministic regime, measured differences in thermodynamic entropy are, with high probability, approximately equal to Gibbs entropy differences. As we will now argue, in that regime Boltzmann entropy differences are approximately equal to Gibbs entropy differences. Therefore, in the quasi-deterministic regime, Boltzmann entropy differences are approximately equal to thermodynamic entropy differences.[15]

Consider some Boltzmannian energy shell Γ_E. Let $\Gamma_E(\boldsymbol{\lambda}_a)$ and $\Gamma_E(\boldsymbol{\lambda}_b)$ be the subsets of Γ_E that are compatible with those values of the external parameters. In the quasi-deterministic regime, we expect each $\Gamma_E(\boldsymbol{\lambda})$ to be dominated by a single macrostate, $M_E(\boldsymbol{\lambda})$, which takes up the vast majority of its Liouville measure. Therefore, in the quasi-deterministic regime,

$$\begin{aligned} S_B(\boldsymbol{\lambda}_b) - S_B(\boldsymbol{\lambda}_a) &= k \log \left(\mu(M_E(\boldsymbol{\lambda}_b)/\mu(M_E(\boldsymbol{\lambda}_a)) \right) \\ &\approx k \log \left(\mu(\Gamma_E(\boldsymbol{\lambda}_b)/\mu(\Gamma_E(\boldsymbol{\lambda}_a)) \right) \end{aligned} \tag{21}$$

In the quasi-deterministic regime, a canonical distribution is sharply peaked near its expectation value. If $\tau_{\beta, \boldsymbol{\lambda}_a}$ and $\tau_{\beta, \boldsymbol{\lambda}_b}$ are both peaked around a common energy expectation value E, with the same spread in energy, and if Γ_E is an energy shell focused on E, with a width comparable to the spread of energy of those canonical distributions,

$$S[\tau_{\beta, \boldsymbol{\lambda}_b}] - S[\tau_{\beta, \boldsymbol{\lambda}_a}] \approx k \log \left(\mu(\Gamma_E(\boldsymbol{\lambda}_b))/\mu(\Gamma_E(\boldsymbol{\lambda}_a)) \right). \tag{22}$$

Therefore, in this regime, the difference of the Boltzmann entropies $S_B(\boldsymbol{\lambda}_b)$ and $S_B(\boldsymbol{\lambda}_a)$ will approximate the difference of thermodynamic entropy of the corresponding thermodynamic states.

6 Foundations and keystones: on the relation between "Boltzmannian" and "Gibbsian" methods

It has become a commonplace of the philosophical literature on statistical mechanics that there are rival approaches to statistical mechanics, usually

[15] See Werndl and Frigg (2020b) for a detailed discussion of conditions under which Boltzmann entropy differences agree with Gibbs energy differences.

called “Boltzmannian” and “Gibbsian.” The terminology is unfortunate, as both have their roots in the work of Boltzmann, and, in particular, the use of an imaginary ensemble of systems, usually thought to be a hallmark of the “Gibbsian” approach, was originated by Boltzmann. Sometimes, the contrast is expressed as one between “individualist” and “ensemblist” approaches (see Goldstein 2019). This is also misleading, as the point of introducing ensembles is to consider the consequences of an incomplete specification of initial conditions for the behaviour of an individual system.

The idea that the families of methods labelled “Boltzmannian” and “Gibbsian” constitute incompatible and competing approaches is alien to the way the founders of the subject thought of it, and to the textbook tradition that emerged.[16] Though Boltzmann’s work on statistical mechanics, which involves a bewildering variety of approaches, might give the impression that he is merely opportunistic and systematic, a careful reading, undertaken by Olivier Darrigol, demonstrates that Boltzmann was deeply concerned with the relations between the various techniques he introduced (Darrigol, 2018). Gibbs himself presented his work, not as a rival to Boltzmann’s, but as extending and building upon it.

Another attitude that has been expressed in the literature is that “Boltzmannian” statistical mechanics forms a foundation for the subject, and “Gibbsian” methods are to be thought of as calculational tools, and nothing more (see Goldstein et al. 2020; Werndl and Frigg 2020a).

Einstein, who, independently of Gibbs, developed much of the machinery of “Gibbsian” statistical mechanics, saw the relation of that machinery to the work of Boltzmann differently. In a letter to Marcel Grossman of September 1901, he wrote,[17]

> Lately I have been engrossed in Boltzmann’s works on the kinetic theory of gases and these last few days I wrote a short paper

[16] Here’s a way to think of it, that I find helpful. Is Boltzmann’s *Lectures on Gas Theory* (Boltzmann 1896, 1898, 1964) a work of “Boltzmannian” or “Gibbsian” statistical mechanics? The H-theorem, usually awarded to the “Boltzmannian” camp, is there, in Boltzmann’s fullest presentation. But so are the ensembles of systems (which Boltzmann called *Ergoden*); see especially sections 26 and 32 of Part II. Is it possible to divide these lectures into two treatises, a “Boltzmannian” and “Gibbsian” one, each capable of standing on its own?

[17] Ich habe mich in letzer Zeit gründlich mit Boltzmanns Arbeiten über kinetische Gastheorie befaßt & in den letzen Tagen selbst eine kleine Arbeit geschreiben, welche den Schlußstein einer von ihm begonnen Beweiskette liefert.

> myself that provides the keystone in the chain of proofs that he had started (Einstein 1987b, p. 181, from Einstein 1987a, p. 315).

We have here two architectural metaphors, *foundation* and *keystone*. The two metaphors convey very different conceptions of the relation between "Boltzmannian" and "Gibbsian" methods in statistical mechanics. The relation of a foundation to the structure built on it is one of asymmetric dependence. The foundation does not need the superstructure for its stability, but the superstructure cannot stand without the foundation. In contrast, an arch is not stable until the keystone is in place, and must be held up by temporary scaffolding. Once the arch is completed, no part of it can stand on its own, without the other parts. It is, of course, possible that Einstein was simply wrong about the relation of his work (and Gibbs') to Boltzmann's. But I invite the reader to seriously consider the possibility that Einstein and Gibbs, both of them deeply immersed in study of Boltzmann's work on statistical mechanics, correctly understood the relation of their work to that of their predecessor.

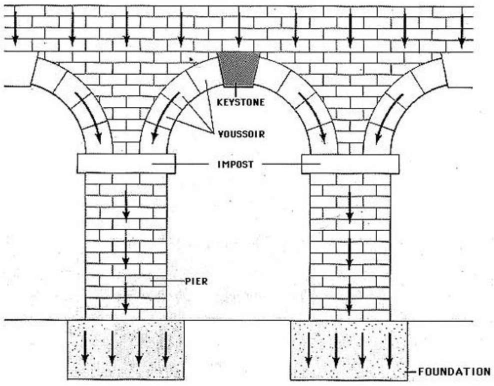

Figure 1: Parts of an arch.

7 Appendix. Proofs

As we have seen, the theorems invoked in this note come in two version, quantum and classical. In this appendix, we will adopt systematically ambiguous notation, and use the notation ρ for either a probability distribution on a classical phase space, represented by density function f_ρ with respect to Liouville measure, or a quantum state, represented by a density operator $\hat{\rho}$ on a Hilbert space. We will use the notation $S[\rho]$ for the Gibbs entropy (classical), or the von Neumann entropy (quantum). We will occasionally use $S(t)$ as shorthand for S applied to the probability distribution (or quantum state) for time t. By *Hamiltonian* evolution we will mean, in the classical context, evolution of a system according to Hamilton's equations, for some (possibly time-varying) Hamiltonian H, and, in the classical context, evolution of a quantum state according to the Schrödinger equation, for some (possibly time-varying) Hamiltonian operator $\hat{H}$. The key feature of Hamiltonian evolution that we need is that, in the classical context, it conserves Liouville measure, and, in the quantum context, it induces a unitary mapping of the Hilbert space to itself.

As mentioned in the main text, for a Hamiltonian $H_{\boldsymbol{\lambda}}$, which may depend on exogenous parameters $\boldsymbol{\lambda}$, a canonical distribution is one that has a density function, with respect to Liouville measure μ,

$$\tau_{\beta,\boldsymbol{\lambda}}(x) = Z^{-1}_{\beta,\boldsymbol{\lambda}}\, e^{-\beta H_{\boldsymbol{\lambda}}(x)}. \tag{23}$$

$Z_{\beta,\boldsymbol{\lambda}}$ is a normalization constant, required to make the integral of $\tau_{\beta,\boldsymbol{\lambda}}$ over all of the accessible region of phase space equal to unity.

$$Z_{\beta,\boldsymbol{\lambda}} = \int e^{-\beta H_{\boldsymbol{\lambda}}}\, d\mu. \tag{24}$$

It depends on the parameter β and on the form of the Hamiltonian $H_{\boldsymbol{\lambda}}$. In fact, this dependence is informative about certain properties of the canonical distribution. In particular, the dependence of $Z_{\beta,\boldsymbol{\lambda}}$ on β yields information about the expectation value of energy, and its variance, for a canonical distribution.

$$\langle H_{\boldsymbol{\lambda}} \rangle_{\tau_{\beta,\boldsymbol{\lambda}}} = -\frac{\partial}{\partial \beta} \log Z_{\beta,\boldsymbol{\lambda}}; \tag{25}$$

$$\mathrm{Var}_{\tau_{\beta,\boldsymbol{\lambda}}}(H_{\boldsymbol{\lambda}}) = \frac{\partial^2}{\partial \beta^2} \log Z_{\beta,\boldsymbol{\lambda}} = -\frac{\partial}{\partial \beta} \langle H_{\boldsymbol{\lambda}} \rangle_{\tau_{\beta,\boldsymbol{\lambda}}}. \tag{26}$$

Because the variance of H is always positive, for any canonical distribution, it follows that, for fixed $\boldsymbol{\lambda}$, expectation value of energy for the canonical distribution, $\langle H_{\boldsymbol{\lambda}} \rangle_{\tau_{\beta,\boldsymbol{\lambda}}}$, is a monotonic decreasing function of β, that is, a monotonic increasing function of temperature. From this it follows that, for fixed $\boldsymbol{\lambda}$, no two distinct canonical distributions have the same expectation value of energy.

A useful property of canonical distributions is the following.

Lemma 1. *For any $T > 0$, the canonical distribution τ_β, where $\beta = 1/kT$, uniquely minimizes the quantity*

$$\langle H \rangle_\rho - TS[\rho].$$

That is, for any distribution ρ,

$$\langle H \rangle_{\tau_\beta} - TS[\tau_\beta] \leq \langle H \rangle_\rho - TS[\rho],$$

with equality only if $\rho = \tau_\beta$.

To prove Lemma 1, we define the *relative entropy* of two probability distributions. If two distributions have density functions ρ, ω, such that ρ is equal to zero wherever ω is, we define their *relative entropy* (also known as the Kullback-Leibler divergence), as

$$S[\rho \,\|\, \omega] = -k \left(\langle \log \omega \rangle_\rho - \langle \log \rho \rangle_\rho \right). \tag{27}$$

It can be proven (see, *e.g.*, Cover and Thomas 1991, §2.6) that $S[\rho \,\|\, \omega] \geq 0$, with equality only if $\rho = \omega$. From this Lemma 1 follows, taking τ_β as ω.

From Lemma 1, and the fact that $\langle H_{\boldsymbol{\lambda}} \rangle_{\tau_{\beta,\boldsymbol{\lambda}}}$, is a monotonic decreasing function of β, we have,

Lemma 2. *The canonical distribution $\tau_{\beta,\boldsymbol{\lambda}}$ uniquely maximizes S among distributions that agree with it on expectation value of energy. That is, if*

$$\langle H_{\boldsymbol{\lambda}} \rangle_\rho = \langle H_{\boldsymbol{\lambda}} \rangle_{\tau_{\beta,\boldsymbol{\lambda}}},$$

then

$$S[\tau_{\beta,\boldsymbol{\lambda}}] \geq S[\rho],$$

with equality only if $\rho = \tau_{\beta,\boldsymbol{\lambda}}$.

We will also use the following facts about the Gibbs/von Neumann entropy S.

Lemma 3. *$S[\rho]$ is conserved under Hamiltonian evolution.*

Lemma 4. *(*Subadditivity of S*). For any system AB consisting of disjoint subsystems A, B, and any probability distribution over AB,*

$$S[\rho_{AB}] \leq S[\rho_A] + S[\rho_B],$$

with equality only if the subsystems A and B are uncorrelated on ρ.

We are now in a position to prove what we set out to prove.

Proposition 5. *Suppose that, at time t_0, two systems A and B have associated with them uncorrelated canonical distributions with parameters β_a and β_b. During the time interval $[t_0, t_1]$, the joint system consisting of A and B undergoes Hamiltonian evolution according to the time-varying Hamiltonian*

$$H_{AB}(t) = H_A + H_B + V_{AB}(t),$$

where the interaction potential V_{AB} is zero at the endpoints of this interval.

$$V_{AB}(t_0) = V_{AB}(t_1) = 0.$$

No restrictions are placed on V_{AB} other than the condition that no net work is done on the system.

$$\langle H_A + H_B \rangle_{t_0} = \langle H_A + H_B \rangle_{t_1}.$$

Then,

- *If $\beta_a = \beta_b$, $\langle \Delta H_A \rangle = 0$,*
- *if $\beta_a > \beta_b$, $\langle \Delta H_A \rangle \geq 0$, and*
- *if $\beta_a < \beta_b$, $\langle \Delta H_A \rangle \leq 0$.*

Proof. We begin with the case $\beta_a = \beta_b$. Then the initial distribution for the joint system AB is a canonical distribution. Because the evolution from t_0 to t_1 is Hamiltonian, S has the same value at the beginning and end of the evolution. The expectation value of total energy is the same at t_1 as it is at t_0. Since, by Lemma 2, the initial canonical distribution is the unique distribution that has the values for S and $\langle H \rangle$ that it does, the final distribution is the same as the initial canonical distribution.

Now consider the case in which $\beta_a \neq \beta_b$. By Lemma 1,

$$\langle H_A(t_0) \rangle - T_a\, S_A(t_0) \leq \langle H_A(t_1) \rangle - T_a\, S_A(t_1), \tag{28}$$

or,

$$\Delta\langle H_A\rangle \geq T_a\Delta\, S_A. \tag{29}$$

Similarly,

$$\Delta\langle H_B\rangle \geq T_b\Delta\, S_B. \tag{30}$$

Because the subsystems A and B are uncorrelated at t_0,

$$S_{AB}(t_0) = S_A(t_0) + S_B(t_0). \tag{31}$$

By Lemma 4,

$$S_{AB}(t_1) \leq S_A(t_1) + S_B(t_1). \tag{32}$$

Because S is conserved under Hamiltonian evolution,

$$S_{AB}(t_1) = S_{AB}(t_0). \tag{33}$$

Combining (31), (32), and (33), we get,

$$S_A(t_0) + S_B(t_0) \leq S_A(t_1) + S_B(t_1), \tag{34}$$

or,

$$\Delta S_A + \Delta S_B \geq 0. \tag{35}$$

Combining (29), (30), and (35), we get

$$\frac{\Delta\langle H_A\rangle}{T_a} + \frac{\Delta\langle H_B\rangle}{T_b} \geq \Delta S_A + \Delta S_B \geq 0. \tag{36}$$

Because total energy is conserved, $\Delta\langle H_B\rangle = -\Delta\langle H_A\rangle$, and so,

$$\left(\frac{1}{T_a} - \frac{1}{T_b}\right)\Delta\langle H_A\rangle \geq 0, \tag{37}$$

or,

$$(\beta_a - \beta_b)\,\Delta\langle H_A\rangle \geq 0. \tag{38}$$

From this follow the claimed assertions. □

Proposition 6. *Consider a system A that, at time t_0, has a probability distribution $\rho_A(t_0)$. Between t_0 and t_1 it interacts successively with systems $\{B_i, i = 1, \ldots, n\}$, which at t_0 have canonical distributions at temperatures T_i, uncorrelated with A. The joint system consisting of A and the systems $\{B_i\}$ undergoes Hamiltonian evolution in the interval, and the coupling of*

A with the systems $\{B_i\}$ conserves total energy. At t_1 the systems are no longer interacting. Let $\langle Q_i\rangle = -\Delta\langle H_{B_i}\rangle$ be the expectation value of the energy received by A from B_i. Then,

$$\sum_{i=1}^{n} \frac{\langle Q_i\rangle}{T_i} \le S_A(t_1) - S_A(t_0).$$

Proof. By the same reasoning that led to (30), we have, for each B_i,

$$\Delta\langle H_{B_i}\rangle \ge T_i\,\Delta S_{B_i}, \tag{39}$$

or,

$$-\frac{\langle Q_i\rangle}{T_i} \ge \Delta S_{B_i}. \tag{40}$$

From this, it follows that,

$$\Delta S_A - \sum_{i=1}^{n} \frac{\langle Q_i\rangle}{T_i} \ge \Delta S_A + \sum_{i=1}^{n} \Delta S_{B_i}. \tag{41}$$

By the same reasoning that led to (35),

$$\Delta S_A + \sum_{i=1}^{n} \Delta S_{B_i} \ge 0. \tag{42}$$

Combining (41) and (42) gives us,

$$\Delta S_A - \sum_{i=1}^{n} \frac{\langle Q_i\rangle}{T_i} \ge 0, \tag{43}$$

or,

$$\Delta S_A \ge \sum_{i=1}^{n} \frac{\langle Q_i\rangle}{T_i}, \tag{44}$$

which is what was to be proven. □

References

Allori, V. (Ed.) (2020). *Statistical Mechanics and Scientific Explanation.* Singapore: World Scientific.

Boltzmann, L. (1877). Bemerkungen über einige Probleme der mechanische Wärmetheorie. *Sitzungsberichte der Kaiserlichen Akademie der Wissenschaften. Mathematisch-Naturwissenschaftliche Classe 75*, 62–100. Reprinted in Boltzmann (1909, 113–148). English translation of Section II in Boltzmann (1966).

Boltzmann, L. (1896). *Vorlesungen Über Gastheorie. I. Thiel.* Berlin: Verlag von Johann Ambrosius Barth.

Boltzmann, L. (1898). *Vorlesungen Über Gastheorie. II. Thiel.* Berlin: Verlag von Johann Ambrosius Barth.

Boltzmann, L. (1909). *Wissenschaftliche Abhandlungen. II. Band.* Leipzig: J. A. Barth.

Boltzmann, L. (1964). *Lectures on Gas Theory.* Berkeley: University of California Press. Translation of Boltzmann (1896, 1898).

Boltzmann, L. (1966). On the relation of a general mechanical theorem to the second law of thermodynamics. In Brush, ed. (1966), pp. 188–193. English translation of section II of Boltzmann (1877).

Brown, H. R. and J. Uffink (2001). The origins of time-asymmetry in thermodynamics: The minus first law. *Studies in History and Philosophy of Modern Physics 32*, 525–538.

Brush, S. G. (Ed.) (1966). *Kinetic Theory, Volume 2. Irreversible Processes.* Oxford: Pergamon Press.

Cover, T. M. and J. A. Thomas (1991). *Elements of Information Theory.* New York: John Wiley & Sons.

Darrigol, O. (2018). *Atoms, Mechanics, and Probability.* Oxford University Press.

Einstein, A. (1987a). *The Collected Papers of Albert Einstein, Volume 1: The Early Years, 1879-1902.* Princeton: Princeton University Press.

Einstein, A. (1987b). *The Collected Papers of Albert Einstein, Volume 1: The Early Years, 1879-1902 (English Translation Supplement)*. Princeton: Princeton University Press.

Garber, E., S. G. Brush, and C. W. F. Everitt (Eds.) (1995). *Maxwell on Heat and Statistical Mechanics: On "Avoiding All Personal Enquiries" of Molecules*. Bethlehem, Pa: Lehigh University Press.

Gibbs, J. W. (1875). On the equilibrium of heterogeneous substances, part I. *Transactions of the Connecticut Academy of Arts and Sciences 3*, 108–248. Reprinted in Gibbs (1906, 55-184).

Gibbs, J. W. (1906). *The Scientific Papers of J. Willard Gibbs, PhD, LLD*, Volume I. New York: Longmans, Green, and Co.

Goldstein, S. (2019). Individualist and ensemblist approaches to the foundations of statistical mechanics. *The Monist 102*, 435–457.

Goldstein, S., J. L. Lebowitz, R. Tumulka, and N. Zanghì (2020). Gibbs and Boltzmann entropy in classical and quantum mechanics. In Allori, ed. (2020), pp. 519–581.

Hájek, A. and C. Hitchcock (Eds.) (2016). *The Oxford Handbook of Probability and Philosophy*. Oxford: Oxford University Press.

Harman, P. M. (Ed.) (1995). *The Scientific Letters and Papers of James Clerk Maxwell, Volume II: 1862-1873*. Cambridge: Cambridge University Press.

La Caze, A. (2016). Frequentism. In Hájek and Hitchcock, eds. (2016), pp. 341–359.

Maroney, O. (2007). The physical basis of the Gibbs-von Neumann entropy. `arXiv:quant-ph/0701127v2`.

Maxwell, J. C. (1878). Tait's "Thermodynamics". *Nature 17*, 257–259, 278–280. Reprinted in Niven (1890, 660–671).

Myrvold, W. C. (2016). Probabilities in statistical mechanics. In Hájek and Hitchcock, eds. (2016), pp. 573–600.

Myrvold, W. C. (2020a). Explaining thermodynamics: What remains to be done? In Allori, ed. (2020), pp. 113–143.

Myrvold, W. C. (2020b). The science of $\Theta\Delta^{\mathrm{cs}}$. *Foundations of Physics 50*, 1219–1251.

Myrvold, W. C. (2021a). *Beyond Chance and Credence.* Oxford: Oxford University Press.

Myrvold, W. C. (2021b). Shakin' all over: Proving Landauer's principle without neglect of fluctuations. *The British Journal for the Philosophy of Science*. Forthcoming. Available at `http://philsci-archive.pitt.edu/17610/`.

Niven, W. D. (Ed.) (1890). *The Scientific Papers of James Clerk Maxwell*, Volume Two. Cambridge: Cambridge University Press.

Szilard, L. (1925). Über die Ausdehnung der phänomenologischen Thermodynamik auf die Schwankungserscheinungen. *Zeitschrift für Physik 32*, 753–788. English translation in Szilard (1972).

Szilard, L. (1972). On the extension of phenomenological thermodynamics to fluctuation phenomena. In B. T. Feld, G. W. Szilard, and K. R. Winsor (Eds.), *The Collected Works of Leo Szilard: Scientific Papers*, pp. 70–102. Cambridge, MA: The MIT Press.

Wallace, D. (2016). Probability and irreversibility in modern statistical mechanics: Classical and quantum. Available at `http://dornsife.usc.edu/assets/sites/1045/docs/oxfordstatmech.pdf`.

Werndl, C. and R. Frigg (2020a). Taming abundance: On the relation between Boltzmannian and Gibbsian statistical mechanics. In Allori, ed. (2020), pp. 617–646.

Werndl, C. and R. Frigg (2020b). When do Gibbsian phase averages and Boltzmannian equilibrium values agree? *Studies in History and Philosophy of Modern Physics 72*, 46–69.