1
Introduction

> The instantaneous reversal of the motion of every moving particle of a system causes the system to move backwards, each particle along its path and at the same speed as before …
>
> *(Thomson, 1874)*

Until very recently, the foundations of statistical mechanics were far from satisfactory (Evans, Searles, and Williams, 2009a). Textbooks approach the derivation of the canonical distribution in one of two ways. A common approach is to *postulate* a microscopic definition for the entropy and then to show that the standard canonical distribution function can be obtained by maximizing the entropy subject to the constraints that the distribution function should be normalized and that the average energy is constant. The choice of the second constraint is completely subjective due to the fact that, at equilibrium, the average of every phase function is fixed. The choice of the microscopic expression for the entropy is also *ad hoc.* This "derivation" is therefore flawed.

The second approach begins with Boltzmann's *postulate* of equal *a priori* probability in phase space for the microcanonical ensemble and then derives an expression for the most probable distribution of states in a small subsystem within that much larger microcanonical system. A variation of this approach is to simply *postulate* a microscopic expression for the Helmholtz free energy via the partition function.

The so-called Loschmidt paradox, which so puzzled Boltzmann and his contemporaries, remained unresolved for 119 years after it was first raised. All the equations of motion in mechanics (both classical and quantum) and electrodynamics are time-reversal-symmetric. Time reversibility of the classical equations of motion is trivial to demonstrate. Consider Newton's equations of motion for the positions $\mathbf{q}_i$ of N identical particles subject to interatomic forces $\mathbf{F}_i(\mathbf{q}_1, \ldots, \mathbf{q}_N)$:

$$m\frac{d^2\mathbf{q}_i(t)}{dt^2} = \mathbf{F}_i(\mathbf{q}), \quad i = 1, \ldots, N \tag{1.1}$$

As Loschmidt and Kelvin (separately) noticed (Loschmidt, 1876; Thomson, 1874), time reversal $t \to -t$ leaves Eq. (1.1) unaltered since $(-1)^2 = 1$. This means that if $\mathbf{q}(t); -\tau < t < \tau$ is a solution of the equations of motion, then so too is

Fundamentals of Classical Statistical Thermodynamics: Dissipation, Relaxation and Fluctuation Theorems,
First Edition. Denis J. Evans, Debra J. Searles, and Stephen R. Williams.

$\mathbf{q}(-t)$: $-\tau < t < \tau$. Changing the direction of time inverts every velocity – as per Kelvin's quote above.

The Loschmidt Paradox can be stated quite simply. If all the laws of physics are time-reversal-symmetric, how can one prove a time-asymmetric law like the second "Law" of thermodynamics that states that the entropy of the Universe "tends to a maximum" (Clausius, 1865; Clausius, 1872). Although there have been many attempts over the last century to resolve this paradox, the matter was not really settled until the first proof of a fluctuation theorem in 1994 (Evans and Searles, 1994).

A less well-known problem concerns Clausius' inequality itself. In some ways, this is an even more fundamental problem because it concerns thermodynamics rather than statistical mechanics. Clausius' inequality for the heat $Q_{\rm th}$ transferred to a thermal reservoir states that the cyclic integral $\oint dQ_{\rm th}/T \geq 0$. When this inequality is, in fact, an *equality* (the process is *quasi-static*), we have the usual argument that $\int dQ_{\rm th}/T_{\rm th}$ is a state function and represents the change in the equilibrium entropy of the reservoir, $S_{\rm th}$ and $T_{\rm th}$ is the equilibrium thermodynamic temperature of that reservoir or set of reservoirs. Clausius went on to apply his inequality to the system of interest (soi) and thermal reservoir (th). Indeed, in his original papers he does not distinguish between the two systems.

Now comes the difficulty: when we have a strict inequality $\oint dQ/T > 0$, either the system of interest or the reservoir (or both) is (or are) not in true thermodynamic equilibrium (the process is not *quasi-static*). In this case, what is the temperature? Clausius only defined the temperature for quasi-static or equilibrium processes where the entropy is a state function. In the case of a strict inequality, $\int dQ/T$ is *not* a state function. It is path- and/or history-dependent.

For quasi-static processes (only!), the change in equilibrium entropies of two equilibrium states can be obtained by considering $\int dQ_{\rm th}/T_{\rm th}$ for a reversible (i.e., infinitely slow) pathway between the two equilibrium states. However, if the initial or final states are out of equilibrium or if the pathway connecting the two states is irreversible, the entropy that Clausius defined is ill-defined and so too is the temperature: $T \equiv \partial U/\partial S|_V$, where U is the internal energy, S the (undefined) entropy, and V the volume. This means that the Clausius *inequality* $\oint dQ/T > 0$ is without meaning.

Clausius is famous for his declaration:

> The energy of the Universe is constant. The entropy tends to a maximum.
> *(Clausius, 1865, 1872)*

He did not recognize the fact that he only defined the entropy (and temperature) for reversible processes. This particular difficulty was first discussed in the late nineteenth century by Bertrand (1887) and early in the twentieth century by Orr (1904), Orr (1905), Planck (1905), and Buckingham (1905).

"There are three things in Prof. Orr's article (Orr, 1904) which stand out as of particular importance. (1) He says in substance, though with great moderation, that all proofs of the theorem … when the integral is taken round an irreversible

cycle, are rubbish." Buckingham later discusses problems with writing textbooks while being aware at the time, of some of the difficulties mentioned above. Buckingham continues: "The question how a treatise should be written is not so easily answered. … I do not know of a single book which today deserves the title of 'Treatise on Thermodynamics'." He concluded: "We must leave the question of the proper method for a treatise to the future when the difficulties which now beset us may have vanished." (Buckingham, 1905)

In 1905, Planck responded to Orr (Planck, 1905) agreeing with Orr's concerns on the definition of temperature and saying in part that: "If a process takes place so violently that one can no longer define temperature … , then the usual definition of entropy is inapplicable."

These particular difficulties were only exacerbated in 1902 with the publication (and subsequent circulation) of Gibbs' seminal treatise "Elementary Principles in Statistical Mechanics" (Gibbs, 1981). In his treatise, Gibbs showed that the microscopic expression he identified at equilibrium, as the thermodynamic entropy $S_G(t) \equiv -k_B \int d\mathbf{\Gamma} f(\mathbf{\Gamma};t) \ln[f(\mathbf{\Gamma};t)]$, where $f(\mathbf{\Gamma};t)$ is the N-particle phase space distribution function at time t, is in fact a constant of the motion for autonomous Hamiltonian dynamics! If the initial distribution was not the equilibrium distribution, the Gibbs entropy did not, as Clausius claimed, increase in time until it reached its maximum and the system was effectively in equilibrium. For these systems, the Gibbs' entropy is simply a constant independent of time.

After Boltzmann's death, this distressing state of affairs was reviewed without satisfactory resolution by the Ehrenfests in 1911 (Ehrenfest and Ehrenfest, 1990). (Paul Ehrenfest was a student of Boltzmann.) Indeed ,in the Preface to the (English) Translation, Tatiana Ehrenfest confides: "At the time the article was written [1911], most physicists were still under the spell of the derivation by Clausius of the existence of an integrating factor for the … heat … it became clear to me afterwards, that the existence of an integrating factor has to do only with the differentials $dx_1, dx_2, \ldots, dx_n$ of the *equilibrium* [T. Ehrenfest's italics] parameters $dx_1, dx_2, \ldots, dx_n$, and is completely independent of the direction of time … Nevertheless even today [1959] many physicists are still following Clausius, and for them the second law of thermodynamics is still identical with the statement that entropy can only increase."

The Ehrenfests' article did point out that away from equilibrium entropy was problematic and that for autonomous Hamiltonian systems the entropy defined by Gibbs was indeed a constant of the motion. In Ehrenfest and Ehrenfest (1990, p. 54), they agree with Gibbs that, "From Liouville's theorem, Eqs. (26) and (26'), it follows immediately that the quantity σ [i.e., S_G above] … remains exactly constant during the mixing process." They go on to discuss Gibbs' flawed attempts to resolve the paradox by defining a coarse-grained entropy. This quantity's time dependence is determined by the grain size and is thus not an objective property of the physical system of interest.

The theory of the relaxation to equilibrium has also been fraught with difficulties (Evans, Searles, and Williams, 2009a). First, there was no mathematical definition of equilibrium! The first reasonably general approach to this problem is

summarized in the Boltzmann H-theorem. Beginning with the definition of the H-function, Boltzmann proved that the Boltzmann equation for the time evolution of the single particle probability density implies, for uniform ideal gases, a monotonic decrease of the H-function in time (Boltzmann, 1872) – see the review by Lebowitz (1993) for a modern discussion of Boltzmann's ideas.

However, there are at least two problems with Boltzmann's treatment. First, the Boltzmann equation is valid only for an ideal gas. Second, and more problematic, unlike Newton's equations, Hamilton's principle, or the time-dependent Schrödinger equation, the Boltzmann equation itself is *not* time-reversal-symmetric. It is therefore completely unsurprising that the Boltzmann equation predicts a time-irreversible result, namely the Boltzmann H-theorem.

This leads to a second version of the irreversibility paradox (at least for ideal gases): how can the time-irreversible Boltzmann equation, which leads easily to the time irreversible Boltzmann H-theorem, be derived exactly for ideal gases from time-reversible Newton's equations? This issue was also discussed, without resolution, in the Ehrenfest encyclopedia article (Ehrenfest and Ehrenfest, 1990).

Since our new proof of how macroscopic irreversibility arises from time-reversible microscopic dynamics is valid for all densities, we do not need to directly answer this question in this book. We do make the comment, however, that it is thought that in the ideal gas limit, the Boltzmann equation is exact, but its detailed derivation is beyond the scope of this present book.[1)]

The 1930s saw significant progress in ergodic theory with a proof that for a finite, autonomous Hamiltonian system, whose dynamics preserves a *mixing* microcanonical equilibrium distribution (i.e., a distribution that is uniform over the constant energy phase space hypersurface), averages of physical properties must, in the long-time limit, approach those obtained with respect to that equilibrium microcanonical distribution, regardless of the initial distribution (Sinai, 1976). Later in this book we will give a generalization of the ergodic theory proof. We consider finite systems with autonomous dynamics that are mixing with respect to some possibly thermostatted and/or barostatted equilibrium distribution that is also a solution to the dynamics considered. We show that for such systems, at sufficiently long times, averages of physical phase functions will approach, to arbitrary accuracy, the equilibrium averages taken over their mixing equilibrium distributions, irrespective of the initial distribution.

These proofs are, however, not very revealing. They tell us almost nothing of the relaxation process, only that it takes place. Relaxation is inferred rather than elucidated.

We go on to discuss a new set of theorems and results that, when taken together, provide a completely new approach to establishing the foundations of classical statistical thermodynamics and simultaneously resolving all the issues mentioned above. Each of these theorems is consistent with time-reversible,

1) In Chapter 9, we do make some comments on the relationship between Boltzmann's assumption of molecular chaos (stosszahlansatz in German) and the axiom of causality. It is this assumption that breaks time reversal symmetry in the Boltzmann equation.

deterministic dynamics. Indeed, time reversibility of the underlying equations of motion is the key component to proving these theorems. We do comment that there are stochastic and/or quantum versions of some of the theorems. Each of these theorems is exact for systems of arbitrary size: taking the thermodynamic limit is not required. The theorems are valid for arbitrary temperatures and densities. The theorems are exact arbitrarily near to, or far from, equilibrium. Assumptions about being arbitrarily close to equilibrium, so that the response of systems to external forces is linear, are not required. In the process of deriving these theorems, the so-called "Laws" of thermodynamics cease to be unprovable "Laws" and instead become mathematical theorems.

The first step toward understanding how macroscopic irreversibility arises from microscopically time-reversible dynamics came in 1993 when Evans, Cohen, and Morriss (1993) proposed the first so-called fluctuation relation. By generalizing concepts from the theory of unstable periodic orbits in low-dimensional systems, these authors proposed a heuristic, asymptotic argument for the relative probability of seeing sets of trajectories and their conjugate sets of antitrajectories in nonequilibrium steady states maintained at constant internal energy. In the following year, Evans and Searles (1994) published the first mathematical proof of a fluctuation theorem. A generalized and detailed proof of the Evans–Searles fluctuation theorem is given in Chapter 3. This proof concerns the relative probability of fluctuations in sign of a quantity now known as the time-averaged *dissipation function*. Unsurprisingly, fluctuation theorems lead to many new results. This is what the present book sets out to describe. It used to be said that there are very few exact results that are known for nonequilibrium many-body systems. This is no longer the case.

In Chapter 3, we prove the second law inequality (Searles and Evans, 2004), and the nonequilibrium partition identity (Morriss and Evans, 1985; Carberry *et al.*, 2004; Evans and Searles, 1995). These are simple mathematical consequences of the fluctuation theorem. The second law inequality is, in fact, a generalization of the second "Law" of thermodynamics that is valid for finite, even small systems, observed for finite, even short, times. Classical thermodynamics applies to only large, in principle infinite, systems either at equilibrium or in the infinitely slow, or quasi-static, limit.

Dissipation was first explicitly defined in 2000 by Searles and Evans (2000a), although it was, of course, implicit in the earlier proofs of the Evans–Searles fluctuation theorems in 1994, et seq. It is also implicit in many of Lord Kelvin's papers in the late nineteenth century. The dissipation function has many properties, but its original definition directly involved sets of trajectories and their conjugate sets of time-reversed antitrajectories. For classical N-particle systems, the specification of all the coordinates and momenta of all the atoms in the system completely describes the microstate of a classical system. We define the phase space vector $\mathbf{\Gamma} = (\mathbf{q}_1, \ldots, \mathbf{q}_N, \mathbf{p}_1, \ldots, \mathbf{p}_N)$ of the positions $\mathbf{q}_i$ and momenta $\mathbf{p}_i$ of the N particles. We imagine an infinitesimal set of phases inside an infinitesimal volume $dV_{\mathbf{\Gamma}}(\mathbf{\Gamma})$ in phase space. For simplicity, we assume that the system is autonomous (i.e., the

equations of motion for all the particles, $\dot{\mathbf{\Gamma}}(\mathbf{\Gamma}, t)$, do not refer explicitly to time $\dot{\mathbf{\Gamma}}(\mathbf{\Gamma})$; any external fields are time-independent).

As time evolves, this set will trace out an infinitesimal tube in phase space. We follow this tube for a time interval $(0, t)$. At time t, an initial phase space vector $\mathbf{\Gamma}$ has evolved to the position $S^t\mathbf{\Gamma}$, where S^t is the phase space – time evolution operator. If we take the set of phase points inside the infinitesimal volume $dV_{\mathbf{\Gamma}}(S^t\mathbf{\Gamma})$ and reverse all the momenta leaving all the particle positions unchanged, we have the phase vector $M^T S^t\mathbf{\Gamma}$, where M^T is a time-reversal mapping: $M^T(\mathbf{q}, \mathbf{p}) = (\mathbf{q}, -\mathbf{p})$.

If we now imagine following the natural motion of this mapped set forward in time from time t to $2t$, we arrive at the phase point $S^t M^T S^t\mathbf{\Gamma}$. Because the equations of motion are time-reversal-symmetric, the final set of phase points will have the same position coordinates but the opposite momenta to the original set of time zero phases: $S^t M^T S^t\mathbf{\Gamma} = M^T\mathbf{\Gamma}$. This is the fundamental property of time-reversible dynamics discussed in Kelvin's quote at the beginning of this chapter. This time reversibility property is exploited directly in the definition of the dissipation function. We will give a more detailed description of reversibility using a more precise notation in Chapter 2 – especially in Section 2.1.

The time integral of the dissipation function is simply defined as the natural logarithm of the probability ratio of observing at time zero the conjugate sets of trajectories inside phase space volumes $\delta V_{\mathbf{\Gamma}}(\mathbf{\Gamma})$, $\delta V_{\mathbf{\Gamma}}(M^T S^t\mathbf{\Gamma})$:

$$\lim_{\delta V_\Gamma \to 0} \frac{p(\delta V_{\mathbf{\Gamma}}(\mathbf{\Gamma}); 0)}{p(\delta V_{\mathbf{\Gamma}}(M^T S^t\mathbf{\Gamma}); 0)} \equiv \exp\left[\int_0^t ds\,\Omega\,(S^s\mathbf{\Gamma})\right] \tag{1.2}$$

The small phase space volume $\delta V_{\mathbf{\Gamma}}(\mathbf{\Gamma})$ defines an initial set of phase space trajectories. The volume $\delta V_{\mathbf{\Gamma}}(M^T S^t\mathbf{\Gamma})$ defines the conjugate set of the antitrajectories. Going forward in time from $\delta V_{\mathbf{\Gamma}}(M^T S^t\mathbf{\Gamma})$ is like going backward in time from $\delta V_{\mathbf{\Gamma}}(S^t\mathbf{\Gamma})$ except that all the momenta are reversed. For Eq. (1.2) to be well defined requires that the system should be *ergodically consistent*, that is, if the numerator is nonzero for initial phases inside some specified phase space domain D, then the denominator must also be nonzero. This condition ensures that the dissipation function is well defined *everywhere* inside the ostensible phase space domain, D.

As a historical remark, we can see from the definition, Eq. (1.2), that ergodic consistency guarantees the existence of (almost all) conjugate phase space trajectory/antitrajectory pairs. However, the mere existence of these pairs of trajectories by no means implies that the probability ratio of observing infinitesimal *sets* of these conjugate trajectory pairs is unity, as Loschmidt tried to imply. Once you have written down Eq. (1.2) for the relative probability of seeing a set of trajectories and its conjugate set of antitrajectories, it seems obvious that Loschmidt's assertion of both sides of Eq. (1.2) equaling unity is wrong. One must not make the mistake of discussing *individual* conjugate phase space trajectories rather than conjugate *sets* of trajectories. The probability of observing any individual phase space trajectory is precisely zero! Their rephrasing of Eq. (1.2) would have been ill defined, leading to zero divided by zero on the left-hand side.

We will see in Chapter 5 that an equilibrium state is characterized by a set of equations of motion and a phase space distribution for which the dissipation

function is identically zero everywhere in phase space. Thus, for equilibrium states alone the probabilities of observing every infinitesimal set of trajectories and its conjugate set of antitrajectories are identical. Loschmidt's assertion is correct only for equilibrium distributions. Indeed, this gives statistical thermodynamics, for the first time, a mathematical definition of an equilibrium system.

Although the definition of the dissipation function may appear rather abstract and mathematical, it turns out that in the linear regime close to equilibrium the average of the dissipation function is equal to a quantity that is familiar in linear, irreversible thermodynamics, namely the spontaneous entropy production. For systems that are driven by an applied dissipative field (e.g., an electrically conducting system being driven by an electric field), the average dissipation is equal to the average power dissipated in the system divided by the thermodynamic temperature of the surrounding thermal reservoir to which the dissipated work, on average, eventually relaxes. A notable aspect of our exposition is the fact that except at equilibrium, entropy plays no role. This neatly bypasses the objections of Bertrand, Orr, and Buckingham to the Clausius inequality for non-quasi-static processes.

The first theorem that referred to dissipation was the Evans–Searles fluctuation theorem (Evans and Searles, 1994) (FT). This theorem considers systems with time-reversible dynamics where the initial distribution of phases is even in the momenta and which satisfies the condition of ergodic consistency. It states that for such systems the ratio of probabilities that the time-averaged dissipation function $\overline{\Omega}_t$ takes on an arbitrary value in the range $A \pm dA$, compared to the negative of that value $-A \pm dA$ satisfies the following equation

$$\frac{p(\overline{\Omega}_t = A)}{p(\overline{\Omega}_t = -A)} = e^{At} \tag{1.3}$$

where $p(\overline{\Omega}_t = A)$ represents the ratio of probabilities that the time-averaged dissipation function $\overline{\Omega}_t$ takes on an arbitrary value in the range $A \pm dA$. This shows that the probability of positive dissipation is exponentially more likely than negative dissipation and, moreover, the argument of the exponential is extensive in both the number of particles in the system N and the averaging time t. Equation (1.3) has been confirmed both by molecular dynamics computer simulations and in actual laboratory experiments. The first unambiguous laboratory demonstration of a fluctuation relation was conducted in 2002 using a colloidal suspension and optical tweezers (Wang *et al.*, 2002).

A trivial consequence of the FT is the second law inequality, which states that, if we average the response of repeated experiments on our system with macroscopically identical initial conditions, the so-called ensemble average of the time-averaged dissipation $\langle \overline{\Omega}_t \rangle$ is nonnegative:

$$\langle \overline{\Omega}_t \rangle \geq 0, \quad \forall t \tag{1.4}$$

This does not imply that the instantaneous ensemble-averaged dissipation is nonnegative. This ensemble-averaged *instantaneous* dissipation $\langle \Omega(t) \rangle$ may be

positive or negative, but it is, of course, positive more often than it is negative in order to satisfy Eq. (1.4).

The second law inequality also shows that

$$\Omega(\mathbf{\Gamma}) = 0, \quad \forall \mathbf{\Gamma} \in D \Leftrightarrow \langle \overline{\Omega}_t \rangle = 0 \tag{1.5}$$

The proof is rather straightforward. Obviously, the left-hand side implies the right. Does the right imply the left? Suppose the ensemble-averaged time integral of the dissipation is not identically zero everywhere. Average the dissipation over some possibly short time interval $(0,t)$. Ergodic consistency implies the existence of conjugate sets of trajectories with opposite values for the time-averaged dissipation $\pm(A + dA)$. Applying the FT to each conjugate set with time-averaged dissipation $\pm(A + dA)$ shows that positive dissipation is exponentially more likely than negative for each value of $|A|$ that is observed. If we now average over all possible values for $|A|$ for which there is nonzero dissipation, we see that $\langle \Omega_t \rangle > 0$. For any nonequilibrium system, the ensemble average of the time-integrated dissipation must be strictly positive. So, if the dissipation is nonzero anywhere in the allowed phase space and the system is ergodically consistent, then the time-averaged, ensemble-average dissipation must be strictly positive. The only states where the ensemble-averaged, time-averaged dissipation is zero are equilibrium states where the instantaneous dissipation is identically zero everywhere in the allowed phase space.

The recently discovered dissipation theorem (Evans, Searles, and Williams, 2008a,b) (Chapter 4) states that the ensemble average of an arbitrary, integrable function of phase $B(\mathbf{\Gamma})$ is related to the time integrals of the correlation function of the dissipation function with the phase variable:

$$\langle B(t) \rangle = \langle B(0) \rangle + \int_0^t ds \langle B(s) \Omega(0) \rangle \tag{1.6}$$

The dynamics employed for evaluating *all* functions on both sides of Eq. (1.6) employs natural system dynamics including any external fields and/or thermostats. This result is valid arbitrarily far from equilibrium and for systems of arbitrary size. In systems where an externally applied field is responsible for driving the system out of equilibrium in the weak field regime where the response to this field is linear, Eq. (1.6) reduces to the very well known Green–Kubo linear response equations (Evans and Morriss, 1990).

Since the instantaneous average dissipation is zero for equilibrium systems, Eq. (1.6) shows that, in the absence of an external field, ensemble averages of phase function never change for systems at equilibrium. It turns out that for equilibrium systems the equilibrium distribution itself never changes.

Together with the definition of dissipation, a second very important definition is that of an ΩT-mixing system. A system is said to be ΩT-mixing if infinite time integrals of ensemble averages of phase variables $B(\mathbf{\Gamma})$, representing physical observables like pressure, stress, energy, and so on, multiplied by the dissipation function and evaluated at time zero are convergent: $(\lim_{t\to\infty} \left| \int_0^t ds \, \langle B(s) \, \Omega(0) \rangle \right| = \text{const} < \infty)$. A system of harmonic oscillators

with zero friction is obviously not ΩT-mixing. ΩT-mixing is a more physically relevant condition than the mixing condition met in ergodic theory. From Eq. (1.6), we see that, if an autonomous system is ΩT-mixing, then at long times the ensemble average of physical phase functions become time-independent at long times. At long times, ΩT-mixing systems must therefore relax either toward nonequilibrium *steady* states or toward equilibrium states. No other possibilities exist.

If the infinite time integral of ensemble averages of time correlation functions of physical phase functions all $A(\mathbf{\Gamma})$ and $B(\mathbf{\Gamma})$ is finite (i.e., $\lim_{t\to\infty}\left|\int_0^t ds\,\langle A(0)B(s)\rangle\right| = \text{const} < \infty$) when the ensemble average of $A(\mathbf{\Gamma})$ is zero, that is, $\langle A(\mathbf{\Gamma})\rangle = 0$, then the system is termed *T-mixing.* Obviously all T-mixing systems are ΩT-mixing. ΩT-mixing systems are not necessarily T-mixing. Note that any phase function with a nonzero ensemble average (say $\widetilde{A}(\mathbf{\Gamma})$) can be transformed into one with zero average, $\widetilde{A}(\mathbf{\Gamma}) - \langle\widetilde{A}(\mathbf{\Gamma})\rangle = A(\mathbf{\Gamma})$.

The dissipation function, ergodic consistency, and the T-mixing condition hold over some specified phase space domain *D.* For example, while particle momenta may be unbounded, the particle coordinates are usually defined only over a fixed region of the physical space. A system is said to be *physically ergodic* over some specified phase space domain if time averages of phase functions representing physical observables taken along almost any phase space trajectory equal late-time ensemble averages taken over any ensemble of initial states.

T-mixing systems must be physically ergodic over that specified phase space domain. If they were not, we could easily construct time correlation functions of physical observables that would never decay to zero. Any initial static correlation between the phase functions would be preserved forever, thereby violating the condition of T-mixing.

Physically, ergodic systems need not be *ergodic over phase space.* Different initial phase space vectors generate, via their different trajectories, different nonintersecting sets of phase space subdomains – one subdomain corresponding to each phase space trajectory and parameterized by time $(0, \infty)$. If the time average of physical properties along each of the different trajectories is independent of the particular trajectory, the system may be *physically ergodic* but not *ergodic over phase space.* This could occur because each trajectory shadows the other trajectories in a densely woven "mat." In this book, we will deal almost exclusively with *physical ergodicity,* which we will refer to simply as *ergodicity*. On the rare occasions that we refer to *ergodicity over phase space*, we will make that explicit at the time. Of course, if a system is *ergodic over phase space*, it must also be *physically ergodic.*

The equilibrium relaxation theorem (Evans, Searles, and Williams, 2009a,b) derived in Chapter 5 states that autonomous N-particle T-mixing systems that may be isolated or perhaps interact with a heat bath and whose initial distributions are even functions of the momenta will, at sufficiently long times, relax toward a unique equilibrium state and that

$$\lim_{t\to\infty}\langle\Omega(S^t\mathbf{\Gamma})\rangle = 0, \quad \forall\mathbf{\Gamma}\in D \tag{1.7}$$

For various forms of thermostat or ergostat, the unique forms of these equilibrium distributions can be determined explicitly using the various individual forms of the equilibrium relaxation theorem. Since *any* reasonably smooth deviation from the unique equilibrium dissipation causes the ensemble-averaged dissipation to be positive, the only conclusion from Eq. (1.7) is that, in the infinite time limit, the system apparently relaxes to its unique equilibrium distribution.

For constant energy dynamics, the equilibrium distribution is uniform over the energy hypersurface[2)] in phase space. The equilibrium relaxation theorem therefore gives a proof of Boltzmann's postulate of equal *a priori* probability for constant energy systems. The relaxation theorem does not imply that all relaxation processes are monotonic in time (i.e., averages of phase functions change monotonically). This is just as well, since experience shows that most relaxation processes are *not* monotonic. For thermostatted systems where the number of particles and the volume are fixed, the unique equilibrium distribution is the well-known canonical distribution postulated by Boltzmann and Gibbs.

An interesting result that we obtain from the equilibrium relaxation theorems is that relaxation to equilibrium *cannot* take place in finite time. In a sense, the equilibrium *distribution* is never reached. It is only *averages* of physical properties that approach, in the infinite time limit, the values one would obtain from a true equilibrium distribution. The actual time dependent phase space distribution becomes, at long times, ever more tightly folded upon itself. It never *becomes* a smooth equilibrium distribution. However, as the equilibrium relaxation theorems prove, the ensemble-averaged dissipation does go to zero in the infinite time limit and in that infinite time limit the distribution must be the unique smooth equilibrium distribution at least as can be ascertained by computing averages of physical phase functions like the dissipation function.

Having determined the equilibrium distribution for systems in contact with a heat reservoir, we show that the standard expression for the change in the calorimetric entropy of the system of interest, $\Delta S_{\rm soi} = \int dQ_{\rm soi}/T$, where $dQ_{\rm soi}$ is the change in the heat added to the system of interest, is, in fact, for quasi-static processes (processes carried out in the infinitely slow limit) a path- and history-independent state function. We show that the so-called integrating factor for the heat, namely $1/T$, which generates the corresponding state function, is in fact unique. No other integrating factor (e.g., $1/T^3$) can generate a state function from the heat. The integrating factor comes directly from the form of the equilibrium canonical distribution function, which is itself unique.

For macroscopic systems, we also derive the fundamental equation for the first and second "laws" of thermodynamics. This equation relates changes in the internal energy U to the equilibrium temperature T appearing in the equilibrium phase space distribution function, the change in the calorimetric entropy, the mechanical pressure p, and the change in the volume dV:

$$dU = TdS - pdV \tag{1.8}$$

2) The "hypersurface" is defined as $\lim_{\delta E \to 0} \{\mathbf{\Gamma} : E < H(\mathbf{\Gamma}) < E + \delta E\}$.

In Eq. (1.8), all quantities are for the system of interest. This macroscopic result is obtained entirely from microscopic or molecular expressions for the various variables.

We also show the identity (up to an arbitrary additive constant) of the Gibbs entropy and the newly defined irreversible calorimetric entropy. The equivalence of changes in the Gibbs and irreversible calorimetric entropies is valid even for irreversible processes where (and unlike Clausius) we take the temperature at any point in a process to be the equilibrium thermodynamic temperature the system would relax to if it was so allowed. The nonequilibrium temperature is, in fact, the equilibrium thermodynamic temperature of the *underlying equilibrium state* toward which the nonequilibrium system is trying to relax.

The derivation of Eq. (1.8) for quasi-static processes (only) is completely consistent with Tatiana Ehrenfest's statement quoted above that, effectively, Eq. (1.8) is "completely independent of the direction of time" (Ehrenfest and Ehrenfest, 1990).

In Chapter 6, we discuss the steady-state relaxation theorem. For systems that are initially in equilibrium for the zero-field dynamics, if a dissipative field is then applied to the system and it is T-mixing, the system will eventually relax to a physically ergodic, nonequilibrium steady state. At long times, time averages of physical phase functions equal late-time ensemble averages. Further we will show that, if the initial equilibrium distribution is perturbed by some reasonably smooth deviation function (even in the particle momenta), the final steady state is independent of the initial perturbation.

Also in Chapter 6, we discuss asymptotic steady-state fluctuation theorems (Searles and Evans, 2000b; Williams, Searles, and Evans, 2006; Searles, Rondoni, and Evans, 2007). For T-mixing systems, these steady-state fluctuation relations are valid even for large deviations from the mean behavior of the system.

In Chapter 7, we describe more theoretical applications of the fluctuation, dissipation, and relaxation theorems. A proof is given of the zeroth law of thermodynamics (Evans, Williams, and Rondoni, 2012); a discussion is given of heat flow and (Evans, Searles, and Williams, 2010) temperature quenches from the point of view of nonequilibrium statistical mechanics. A discussion is given on the relaxation of a color field gradient in a system where the Hamiltonian is color blind. In the linear response regime, as far as its Hamiltonian can sense, the system is in equilibrium. Finally, we give a derivation of an instantaneous fluctuation theorem (Petersen, Evans, and Williams, 2013).

In Chapter 8, we discuss the Crooks fluctuation relations (Crooks, 1998) and the Jarzynski equality (Jarzynski, 1997). These relations show how equilibrium free energy differences can be computed from nonequilibrium path integrals of the work. Using various generalizations of these relations we give a mathematical proof of Clausius' inequality for thermal reservoirs in contact with our system of interest. We consider a set of large thermal reservoirs at a set of temperatures. Because the reservoirs are large compared to the system of interest, they can be regarded as being in thermodynamic equilibrium. We prove (Evans, Williams, and Searles, 2011) for systems that have a periodic response to some cyclic protocol, the ensemble average of the cyclic time integral of the heat transferred to

the reservoirs divided by the corresponding reservoir temperature is nonnegative. Clausius proved his inequality by *assuming* the second law of thermodynamics – the impossibility of constructing a perpetual motion machine of the second kind. Our proof makes no such assumption. Since Clausius' inequality is often taken as the most fundamental statement of the second law, our proof constitutes a direct proof of this statement of the second "Law." We show that it is true only if the system responds periodically to the cyclic protocol (not all systems do this of course), and it is true only if we take the ensemble-averaged response. A single cycle for an individual system, if it is small, may not satisfy Clausius' inequality as it applies to the reservoir.

We also show that, if the reservoirs are small and cannot be regarded as being in thermodynamic equilibrium, the ensemble average of the cyclic integral for the reservoir still satisfies Clausius' inequality. Of course, it only applies if the system responds periodically. At each point in the cycle, the temperature appearing in our generalization of Clausius' inequality is the equilibrium temperature that the entire system would relax to, if the execution of the protocol is stopped and the entire system is allowed to relax to equilibrium.

An immediate consequence of our proof of Clausius inequality for the reservoir is that the change in the entropy of the "universe": $dQ_{\rm th}/T_{\rm th} + dQ_{\rm soi}/T_{\rm soi} = 0$, where "soi" denotes the system of interest, which, by construction, is in thermal contact with the thermal reservoir "th," and is precisely zero. This result is valid for both quasi-static and nonequilibrium processes far from equilibrium using the irreversible calorimetric definition of the entropy. Since we have already proved the equivalence of changes in the irreversible calorimetric and Gibbs entropies, this new result is consistent with the observation made by Gibbs that the Gibbs entropy for an autonomous Hamiltonian system is a constant of the motion. This, of course, contradicts the claim by Clausius that the entropy of the "Universe" tends to a maximum. Furthermore, because we give meaning to temperature far from equilibrium, unlike Clausius' original inequality our result is well defined away from equilibrium and is immune to the criticisms made by Bertrand (1887), Orr (1904), and Buckingham (1905) of the original Clausius *inequality*.

Entropy and dissipation are thus seen to be completely complementary. Away from equilibrium, dissipation is the function that is central to all the theoretical results while entropy plays only a trivial roll. At equilibrium or in the quasi-static limit, dissipation is zero *by definition*, while entropy is one of the key quantities in equilibrium thermodynamics.

In Chapter 9, we revisit the proof of the Evans–Searles FT and discuss the role played therein by the axiom of causality (Evans and Searles, 1996). We prove that in an anti-causal Universe there is an anti-second "Law" of thermodynamics and that ultimately the explanation for the macroscopic irreversibility we see around us is causality. In very few discussions of irreversibility is it realized that, if you apply a time-reversal mapping to a system trajectory, not only do you reverse the direction of the flow of heat and work but the causal response to some time-dependent field becomes anti-causal! Fluxes respond to changes in field strength *before* those changes occur!

If we watch a movie played backwards of macroscopic machines in motion, not only will we see examples of "perpetual motion machines of the second kind" but

we will also see a Universe where effect *precedes* the cause. The transient response to a sudden application of a cause will have the opposite sign to that observed in the forward movie, but that transient response will start *before* the change in the cause has actually occurred!

For example, in a viscometer that is loaded with a viscoelastic fluid, the shear stress in an anti-causal Universe not only has the opposite sign to that which it has in our causal Universe but it will begin to respond (negatively) *before* a shear rate is applied. Likewise, it will begin to decrease towards zero before, not after, the strain rate has been set to zero!

In a causal Universe, one needs to compute the probabilities of events occurring at a time t from the probabilities of prior events and not from the probabilities of events at times later than t. This assumption of causality breaks the time reversal symmetry of the whole system while retaining time-reversible equations of motion.

The assumption of causality seems so ingrained and natural to the human way of thinking that we often do not realize that it is an assumption. It is this assumption, or rather it is the use of this axiom in the proofs of the Evans–Searles and Crooks FTs, that breaks the symmetry of time and leads to the second law inequality rather than an anti-second law inequality.

The principle of least action, which is completely time-reversal-symmetric, does not contain sufficient information to prove any fluctuation theorem. The equations of motion of mechanics must be supplemented with the axiom of causality to predict the operation of machines, engines, and devices in the real world. The axiom is constantly being applied without us even noticing, precisely because it seems so natural. The response of a system (engine) at a given time is obtained by convolving the response function for the system with the time-dependent driving force backward over the past history and *not* over its future. The underlying equations of motion themselves retain their time reversal symmetry.

A clear example of the unrecognized application of the axiom of causality is in the Mori–Zwanzig projection operator formalism – see Zwanzig (2001, Chapter 8). This formalism leads in the linear response limit, to an exact reformulation of the response of a system to time-dependent dissipative fields in the form of a frequency- and wave vector-dependent generalized Langevin equation. In the time domain, the memory kernel associated with the generalized friction coefficient is convolved *backward* in time with the time-dependent driving force. This breaks the time reversal symmetry inherent in the equations of motion themselves. The temporal convolution is over the half space that describes history rather than the future. The spatial convolution, on the other hand, is over *all* physical space: $\pm\infty$ in each Cartesian dimension.

The axiom of causality is also met in electrodynamics where Maxwell's equations permit two solutions for the vector potential: the advanced and the retarded vector potential. In a well-known textbook, they state with little fanfare: "We can now neglect the term V_2' … for it would make the effect appear before the cause" Corson and Lorrain (1962, p. 445). Panofsky and Phillips (1969) are a little more equivocal on the subject: "but only the minus sign appears to have physical significance"; "the advanced potential … appears to violate elementary notions of causality."

It is interesting to re-examine the Boltzmann equation in the light of these observations. In writing the collision integral in the Boltzmann equation, it is assumed that, *before* collisions of ideal gas atoms, the positions and momenta are uncorrelated. After the collision there is correlation. The collision causes the *post*-collisional correlation. The cause of correlation is the collision, which occurs *before* the effect, which is correlation. In a causal Universe, the cause precedes the effect. This is consistent with the assumption of molecular chaos: *stosszahlansatz*.

If one assumes that the positions and momenta are correlated *before* the collisions, then one forms an anti-Boltzmann equation. This is exactly what one would expect if the Universe was, in fact, anti-causal where the coordinates and momenta *before* the collision are affected by the *later* collision. The effect *precedes* the cause, which is the collision.

So in an anti-causal Universe, dilute gases would be described by this anti-Boltzmann equation and the signs of all the transport coefficients (e.g., shear viscosity or thermal conductivity, etc.) would be opposite to those predicted from the Boltzmann equation. This reversal of signs of the transport coefficients for the anti-Boltzmann equation was first pointed out by Cohen and Berlin (1960). The connection between causality and *stosszahlansatz* is new.

Finally, we argue that in an anti-causal Universe where the future influences the present, the inevitable presence of innately random quantum processes in the future, or indeed the exercise of free will in the future by intelligent beings, makes the present state of the Universe undefined. We argue that the only possible Universe where time increases is, in fact, causal. If time were to decrease rather than increase, an anti-causal Universe would appear identical to our own. So ultimately we live in the only possible Universe and the causal second "Law" behavior is, on average, the only physically possible behavior.

References

Bertrand, J.L.F. (1887) *Thermodynamique*, Gauthier-Villars, Paris.

Boltzmann, L. (1872) Further studies on thermal equilibrium among Gas molecules. *Akad. Wissen. Wien*, **66**, 275.

Buckingham, E. (1905) On certain difficulties which are encountered in the study of thermodynamics. *Philos. Mag.*, **9**, 208.

Carberry, D.M., Williams, S.R., Wang, G.M., Sevick, E.M., and Evans, D.J. (2004) The Kawasaki identity and the fluctuation theorem. *J. Chem. Phys.*, **121**, 8179–8182.

Clausius, R. (1865) Ueber verschiedene Für Die anwendungen bequeme formen Der hauptgleichungen Der mechanischen wärmtheorie. *Ann. Phys. Chem.*, **125**, 353.

Clausius, R. (1872) A contribution to the history of the mechanical theory of heat. *Philos. Mag. J. Sci.*, **43**, 106–115.

Cohen, E.G.D. and Berlin, T.H. (1960) Note on the derivation of the Boltzmann equation from the Liouville equation. *Physica*, **26**, 717.

Corson, D.R. and Lorrain, P. (1962) *Introduction to Electromagnetic Fields and Waves*, W. H. Freeman and Company, San Francisco, CA.

Crooks, G.E. (1998) Nonequilibrium measurements of free energy differences for microscopically reversible Markovian systems. *J. Stat. Phys.*, **90**, 1481–1487.

Ehrenfest, P. and Ehrenfest, T. (1990) *The Conceptual Foundations of the Statistical*

Approach to Statistical Mechanics, Dover, Mineola, NY.

Evans, D.J., Cohen, E.G.D., and Morriss, G.P. (1993) Probability of 2nd Law violations in shearing steady-states. *Phys. Rev. Lett.*, **71**, 2401–2404.

Evans, D.J. and Morriss, G.P. (1990) *Statistical Mechanics of Nonequilibrium Liquids*, Academic Press, London.

Evans, D.J. and Searles, D.J. (1994) Equilibrium microstates which generate second Law violating steady states. *Phys. Rev. E*, **50**, 1645–1648.

Evans, D.J. and Searles, D.J. (1995) Steady states, invariant measures, and response theory. *Phys. Rev. E*, **52**, 5839–5848.

Evans, D.J. and Searles, D.J. (1996) Causality, response theory, and the second Law of thermodynamics. *Phys. Rev. E*, **53**, 5808–5815.

Evans, D.J., Searles, D.J., and Williams, S.R. (2008a) On the fluctuation theorem for the dissipation function and its connection with response theory. *J. Chem. Phys.*, **128**, 014504.

Evans, D.J., Searles, D.J., and Williams, S.R. (2008b) On the fluctuation theorem for the dissipation function and its connection with response theory [*J. Chem. Phys.* (2008) **128** 014504], Erratum.. *J. Chem. Phys.*, **128**, 249901.

Evans, D.J., Searles, D.J., and Williams, S.R. (2009a) Dissipation and the relaxation to equilibrium. *J. Stat. Mech: Theory Exp.*, P07029/1–P07029/11.

Evans, D.J., Searles, D.J., and Williams, S.R. (2009b) A simple mathematical proof of Boltzmann's equal a priori probability hypothesis, in *Diffusion Fundamentals III* (eds C. Chmelik, N. Kanellopoulos, J. Karger, and T. Doros), Leipziger Universitatsverlag, Leipzig.

Evans, D.J., Searles, D.J., and Williams, S.R. (2010) On the probability of violations of Fourier's Law for heat flow in small systems observed for short times. *J. Chem. Phys.*, **132**, 024501-1.

Evans, D.J., Williams, S.R., and Rondoni, L. (2012) A mathematical proof of the zeroth "Law" of thermodynamics and the nonlinear fourier "law" for heat flow. *J. Chem. Phys.*, **137**, 194109.

Evans, D.J., Williams, S.R., and Searles, D.J. (2011) A proof of Clausius' theorem for time reversible deterministic microscopic dynamics. *J. Chem. Phys.*, **134**, 204113-1.

Gibbs, J.W. (1981) *Elementary Principles in Statistical Mechanics*, Ox Bow Press, Woodbridge, CT.

Jarzynski, C. (1997) Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.*, **78**, 2690–2693.

Lebowitz, J.L. (1993) Macroscopic laws, microscopic dynamics, Time's arrow and Boltzmann's entropy. *Physica A*, **194**, 1.

Loschmidt, J. (1876) Űber Den zustand Des wärmegleichgewichtes eines systems Von kőrpern Mit rücksicht Auf Die schwerkraft I. *Sitzungsber. Kais. Akad. Wien. Math. Naturwiss. II*, **73**, 128–142.

Morriss, G.P. and Evans, D.J. (1985) Isothermal response theory. *Mol. Phys.*, **54**, 629–636.

Orr, W.M. (1904) On Clausius's theorem for irreversible cycles, and on the increase of entropy. *Philos. Mag. Ser.*, **6** (8), 509.

Orr, W.M. (1905) On Clausius' theorem for irreversible cycles, and on the increase of entropy. *Philos. Mag.*, **9**, 728.

Panofsky, W.K.H. and Phillips, M. (1969) *Classical Electricity and Magnetism*, Addison-Wesley, Reading, MA.

Petersen, C.F., Evans, D.J., and Williams, S.R. (2013) The instantaneous fluctuation theorem. *J. Chem. Phys.*, **139**, 184106.

Planck, M. (1905) On Clausius' theorem for irreversible cycles, and on the increase of entropy. *Philos. Mag.*, **9**, 167.

Searles, D.J. and Evans, D.J. (2000a) Ensemble dependence of the transient fluctuation theorem. *J. Chem. Phys.*, **113**, 3503–3509.

Searles, D.J. and Evans, D.J. (2000b) The fluctuation theorem and green-Kubo relations. *J. Chem. Phys.*, **112**, 9727–9735.

Searles, D.J. and Evans, D.J. (2004) Fluctuations relations for nonequilibrium systems. *Aust. J. Chem.*, **57**, 1119–1123.

Searles, D.J., Rondoni, L., and Evans, D.J. (2007) The steady state fluctuation relation for the dissipation function. *J. Stat. Phys.*, **128**, 1337–1363.

Sinai, Y.G. (1976) *Introduction to Ergodic Theory*, Princeton University Press, Princeton, NJ.

Thomson, W. (1874) Kinetic theory of the dissipation of energy. *Nature*, **9**, 441.

Wang, G.M., Sevick, E.M., Mittag, E., Searles, D.J., and Evans, D.J. (2002) Experimental

demonstration of violations of the second Law of thermodynamics for small systems and short time scales. *Phys. Rev. Lett.*, **89**, 050601.

Williams, S.R., Searles, D.J., and Evans, D.J. (2006) Numerical study of the steady state fluctuation relations Far from equilibrium. *J. Chem. Phys.*, **124**, 194102-1.

Zwanzig, R. (2001) *Nonequilibrium Statistical Mechanics*, Oxford University Press, Oxford.