PHYSICAL REVIEW A VOLUME 18, NUMBER 5 NOVEMBER 1978

Alternative derivation of the classical second law of thermodynamics

I. M. Bassett
School of Physics, University of Sydney, New South Wales, Australia 2006
(Received 3 June 1977)

One of the classical statements of the second law of thermodynamics is derived from an assumption about the behavior of an isolated heat reservoir, together with a further assumption, which seems to be a rather weak one, about initial states. The derivation is otherwise based on the unitary property of time evolution in quantum mechanics. An analogous derivation, within the framework of classical mechanics, is also given; in this case Liouville's theorem substitutes for unitarity.

I. INTRODUCTION

A fundamental explanation of the second law of thermodynamics is often sought in the form of a proof that the entropy (somehow defined in statistical mechanical terms) of a thermally isolated system increases monotonically with time.[1] Without using the notion of entropy, and without establishing any such monotonic approach to equilibrium, it is nevertheless possible to obtain one of the classical statements of the second law from a narrower basis, the main part of which is the proposition that the "heat reservoir" of the classical statement is a physical system which, if isolated, attains an equilibrium state describable by a canonical ensemble. The point of view of this paper is not to seek an "explanation of irreversibility" but rather to dispose, in as convincing a way as possible, of the possibility of "perpetual motion of the second kind."

II. INEQUALITY

The argument is based on the following inequality: If the density matrix $\rho(t)$ of a system at time t is related to that at time zero by a unitary similarity transformation, then

$$\mathrm{Tr}[\rho(t)F(\rho(0))] \geq \mathrm{Tr}[\rho(0)F(\rho(0))], \tag{1}$$

provided both sides of (1) converge and $F(x)$ is a monotonic nonincreasing function of its argument x, i.e.,

$$F(x') \leq F(x) \quad \text{if } x' \geq x. \tag{2}$$

Strictly speaking, the following derivation of (1) assumes that there is a discrete representation in which $\rho(0)$ is diagonal. While it is presumably possible to extend the argument to cover the case of a continuous or partly continuous spectrum, it is not necessary to do so for the purpose of this paper. The representations actually employed for $\rho(0)$ may be labeled by the eigenvalues of energy (together with sets of particle numbers, in Sec. IV). Since the systems dealt with are finite, these representations may be taken to be discrete.

The inequality (1), given (2), may be deduced from the following theorem of real algebra, proved in Appendix A. Let

$$x_1 \leq x_2 \leq x_3 \leq \cdots \tag{3}$$

and

$$y_1 \geq y_2 \geq y_3 \geq \cdots \geq 0. \tag{4}$$

Let $\sum_{i=1}^n y_i$, $\sum_{i=1}^n \alpha_{ij}$, $\sum_{j=1}^n \alpha_{ij}$, and $\sum_{j=1}^n \alpha_{ij} y_j$ converge as $n \to \infty$ and let

$$\alpha_{ij} \geq 0, \quad i,j = 1, 2, \ldots, \tag{5}$$

$$\sum_{i=1}^{\infty} \alpha_{ij} = 1, \quad j = 1, 2, \ldots, \tag{6}$$

$$\sum_{j=1}^{\infty} \alpha_{ij} = 1, \quad i = 1, 2, \ldots. \tag{7}$$

If y_i' is defined by

$$y_i' = \sum_{j=1}^{\infty} \alpha_{ij} y_j, \quad i = 1, 2, \ldots \tag{8}$$

then, provided both sides of the inequality converge,

$$\sum_{i=1}^{\infty} y_i' x_i \geq \sum_{i=1}^{\infty} y_i x_i. \tag{9}$$

To bring (1) into the form of (9), first write (1) in a representation in which $\rho(0)$ is diagonal. Labeling rows and columns of this representation by i and j, the left-hand side of (1)

$$\mathrm{Tr}[\rho(t)F(\rho(0))] = \sum_i \rho_{ii}(t) F(\rho_{ii}(0))$$
$$= \sum_{ij} |U_{ji}|^2 \rho_{jj}(0) F(\rho_{ii}(0)), \tag{10}$$

where U_{ji} is the j,i element of the unitary matrix U which relates $\rho(t)$ to $\rho(0)$,

$$\rho(t) = U^\dagger(t)\rho(0)U(t). \tag{11}$$

The right-hand side of (1)

$$\mathrm{Tr}[\rho(0)F(\rho(0))] = \sum_i \rho_{ii}(0) F(\rho_{ii}(0)). \tag{12}$$

We may assume without loss of generality that the $\rho_{ii}(0)$ are so ordered that

$$\rho_{11}(0) \geq \rho_{22}(0) \geq \rho_{33}(0) \geq \cdots . \quad (13)$$

Setting $y_i = \rho_{ii}(0)$, $x_i = F(\rho_{ii}(0))$, $\alpha_{ij} = |U_{ji}|^2$, and $y_i' = \rho_{ii}(t)$, (1) assumes the form of (9), and the premises of (9), expressed by (3)–(8) together with the convergence conditions preceding (5), are satisfied.

III. SIMPLE APPLICATION OF THE INEQUALITY

A simple physical application of the inequality (1) is obtained as follows. If H_0 is the Hamiltonian of a system at time $t=0$, let the initial density matrix $\rho(0)$ satisfy $\rho(0)=f(H_0)$, where $f(x)$ is a monotonic nonincreasing function of x. It follows that $H_0 = F(\rho(0))$, where $F(x)$ is a monotonic nonincreasing function of x. Inserting this expression for $F(\rho(0))$ into (1) gives $\mathrm{Tr}[\rho(t)H_0] \geq \mathrm{Tr}[\rho(0)H_0]$. Thus, given the constraint on the initial density matrix, an arbitrary time-dependent perturbation to the system Hamiltonian (which is finally switched off) cannot decrease the energy of the system.

[The monotonic initial condition $\rho(0)=f(H_0)$ is a very special one: indeed, although it is, strictly, more general than the canonical equilibrium condition, defined by $f(x)=e^{-\beta x}$, the two conditions are effectively equivalent for ordinary macroscopic systems, as is implied by the result of Appendix B.]

The inequality (1) is valid also if ρ is a finite $N\times N$ matrix. The algebraic theorem required for the finite case can be simply obtained by adding to the conditions (3)–(8) the condition that x_i, y_i, and α_{ij} vanish if i or j exceeds N. A physical example is provided by a single spin 1 in a magnetic field, assumed to be initially describable by a density matrix diagonal and monotonically nonincreasing in the energy; this system cannot have its (mean) energy decreased by any time-dependent perturbation which is finally switched off.

IV. DERIVATION OF THE SECOND LAW

The classical second law of thermodynamics may be expressed in the following form[2,3]: It is impossible to construct an engine which will work in a complete cycle and produce no effect except the external performance of work and the cooling of a heat reservoir.

We assume (in accordance with the remarks in Sec. I) that the heat reservoir has been prepared so that initially, i.e., when first brought into thermal contact with other bodies, it is describable by a canonical ensemble. The following additional assumption will be made for the time being: the "engine" (from now on called the "system proper") is assumed to be initially in "piecemeal equilibrium," i.e., to consist initially of a number of statistically independent parts, each describable by a density matrix of the grand canonical form; the temperature and the chemical potentials can be different in each part. Indicating the νth part of the system proper by the superscript ν, the initial density matrix of the νth part is given by

$$\rho^{(\nu)}(0) = A^{(\nu)} \exp\left[-\beta^{(\nu)}\left(H^{(\nu)} - \sum_i \mu_i^{(\nu)} N_i^{(\nu)}\right)\right], \quad \nu = 1, 2, \ldots . \quad (14)$$

$\beta^{(\nu)}$ is a positive constant, $H^{(\nu)}$ is the Hamiltonian of the νth part, $N_i^{(\nu)}$ the number of molecules of the ith chemical species in the νth part, and $\mu_i^{(\nu)}$ the corresponding chemical potential. $A^{(\nu)}$ is a normalizing constant.

The initial density matrix of the system proper is the direct product of the initial density matrices (14) for the parts. The heat reservoir is assumed to be initially statistically independent of the system proper. Indicating the heat reservoir by the superscript h, the initial density matrix of the heat reservoir is

$$\rho^{(h)}(0) = A^{(h)} \exp(-\beta^{(h)} H^{(h)}) . \quad (15)$$

The initial density matrix $\rho(0)$ of the system consisting of the system proper and the heat reservoir is the direct product

$$\rho(0) = \rho^{(h)}(0) \times \rho^{(1)}(0) \times \rho^{(2)}(0) \times \cdots . \quad (16)$$

$H^{(h)}$ and the $H^{(\nu)}$ all commute with one another since they operate on different spaces. Accordingly

$$-\ln\rho(0) = \beta^{(h)} H^{(h)} + \sum_\nu \beta^{(\nu)}\left(H^{(\nu)} - \sum_i \mu_i^{(\nu)} N_i^{(\nu)}\right) + \text{const} . \quad (17)$$

Hence, taking $F(x) = -\ln x$, the inequality (1) implies

$$\mathrm{Tr}[\rho(t)\beta^{(h)}H^{(h)}] + \sum_\nu \mathrm{Tr}\left[\rho(t)\beta^{(\nu)}\left(H^{(\nu)} - \sum_i \mu_i^{(\nu)} N_i^{(\nu)}\right)\right] \geq \mathrm{Tr}[\rho(0)\beta^{(h)}H^{(h)}] + \sum_\nu \mathrm{Tr}\left[\rho(0)\beta^{(\nu)}\left(H^{(\nu)} - \sum_i \mu_i^{(\nu)} N_i^{(\nu)}\right)\right] . \quad (18)$$

In this expression the quantities $\beta^{(h)}$, $\beta^{(\nu)}$, and $\mu_i^{(\nu)}$ are constants determined by the initial state. Let us suppose that at some time $t=t_1$ the Hamiltonian of each part of the system proper has resumed its initial form, so that the quantity $\mathrm{Tr}[\rho(t)H^{(\nu)}]$ represents the (mean) energy of the νth part. If, for $t=t_1$ the mean energies and mean particle numbers for each part have returned to their initial values, i.e., if

$$\mathrm{Tr}[\rho(t_1)H^{(\nu)}]=\mathrm{Tr}[\rho(0)H^{(\nu)}],$$
$$\mathrm{Tr}[\rho(t_1)N_i^{(\nu)}]=\mathrm{Tr}[\rho(0)N_i^{(\nu)}],$$
$$i=1,2,\ldots,\quad \nu=1,2,\ldots \tag{19}$$

then

$$\mathrm{Tr}[\rho(t_1)H^{(h)}]\geq \mathrm{Tr}[\rho(0)H^{(h)}], \tag{20}$$

i.e., the energy of the heat reservoir cannot have decreased.

The inequality (20) expresses the second law in this case. The scope of (20) is wider than it appears to be at first sight: the initial condition (14) on the system proper may be weakened, and a weaker meaning than (19) can be given to the "cyclic" nature of the process undergone by the system proper.

Consider the following process, in a certain weak sense cyclic. In each time interval $n\tau \leq t \leq (n+1)\tau$, $n=0,1,2,\ldots$, an amount of energy $\geq W>0$, where W is independent of n, is withdrawn from a heat reservoir, a new (identically prepared) heat reservoir being employed in each cycle. The possibility of continuing this process indefinitely can be ruled out provided the initial state of the system proper can be constructed (by the performance of mechanical work) from materials which, at some earlier time $t'<0$ were in "piecemeal equilibrium." Let t'' be the time at which the last cycle ends. Let the inequality (18) be formed for t' as initial time, and $t=t''$ as final time. It may be taken as an implication of the cycle nature of the process undergone by the system proper that the changes in mean energies and particle numbers between t' and t'' are bounded as the number of cycles increases; that is, that the net contribution to the inequality of the terms in $N_i^{(\nu)}$ and $H^{(\nu)}$ is bounded. If the number of cycles could be taken arbitrarily large, the terms in $H^{(h)}$ would become dominant, and the sign of the inequality would be violated.

V. ROLE OF UNITARITY

The classical second law is a statement about the sign of the energy change of a heat reservoir. This sign has been obtained by the present argument as a consequence of certain initial conditions, plus unitarity, Eq. (11). Unitarity follows from the ascription to the system as a whole (including the heat reservoirs) of a Hermitian Hamiltonian (in general time dependent). This is an idealization, since it involves neglecting the thermal interactions which even the best insulated system must have with its surroundings. While such interactions may have a significant effect on the "fine grained" entropy, $-k\,\mathrm{Tr}(\rho\ln\rho)$,[4] they do not have a significant effect on macroscopic energies, which are the quantities in question here. Such energies are not sensitive to the physical conditions at the boundary of a properly insulated vessel, as Joule's experiments incidentally demonstrated.

VI. DERIVATION WITHIN CLASSICAL MECHANICS

The classical version of the inequality (1) is

$$\int d\xi\,\rho(\xi,t)F(\rho(\xi,0))\geq\int d\xi\,\rho(\xi,0)\,F(\rho(\xi,0)). \tag{21}$$

ρ is the classical distribution function, ξ stands for a set of canonical coordinates and momenta, and the range of integration R is the whole physically accessible range of ξ, the whole of "phase space," assumed to be the same at time t and at time zero.

If the system is initially in some state ξ' in R it will be in some definite state ξ'' in R at time t. Conversely (since the equations of motion can be integrated backwards in time as well as forwards), any state ξ'' in R at time t arises from a definite initial state ξ' in R. The system in the state ξ may be though of as a "particle" at the "point" ξ. An ensemble of systems may be represented by a cloud of particles of some density $\sigma(\xi)$ per unit volume. By Liouville's theorem,[5] the correspondence $\xi'\leftrightarrow\xi''$ preserves this density, $\sigma(\xi')=\sigma(\xi'')$. Hence, if the cloud initially fills the whole of phase space at uniform density σ it must do so also at time t. Accordingly, in this case the number of particles in any fixed region of phase space is the same at times t and zero; that is, the number of particles which have left between times zero and t is equal to the number which have entered. Let the whole of phase space be divided into regions $1,2,3,\ldots$ of equal volume V. Let the fraction of the particles initially in the ith region which have moved to the jth at time t be written P_{ij}. Then the number of particles which have moved out of the ith region is $\sigma V\sum_{j(j\neq i)}P_{ij}$; the number of particles which have moved in is $\sigma V\sum_{j(j\neq i)}P_{ji}$. Hence,

$$\sum_{j(j\neq i)}P_{ij}=\sum_{j(j\neq i)}P_{ij},\quad i=1,2,3,\ldots. \tag{22}$$

From the definition of the P_{ij},

$$\sum_{j=1}^{\infty} P_{ij} = 1, \quad i = 1, 2, 3, \ldots . \tag{23}$$

It then follows from (22) that

$$\sum_{i=1}^{\infty} P_{ij} = 1, \quad j = 1, 2, 3, \ldots . \tag{24}$$

Consider a given way of subdividing the phase space into regions of equal volume V, and, associated with this subdivision, an initial distribution function $\rho(\xi, 0)$ which is uniform in each region, i.e., for which

$$\int_i d\xi' \rho(\xi', 0) = V\rho(\xi, 0), \quad i = 1, 2, 3, \ldots , \tag{25}$$

where the integration is over the ith region and ξ is any point in the ith region. If $\mathfrak{N}$ is the total number of particles in the ensemble, the number of particles initially in the ith region is $\mathfrak{N}\int_i d\xi\, \rho(\xi, 0)$ and the number of particles in the ith region at time t is $\mathfrak{N}\int_i d\xi\, \rho(\xi, t)$, where $\rho(\xi, t)$ is the distribution function which follows at time t from the chosen $\rho(\xi, 0)$. Let us write

$$\rho_i(t) = \int_i d\xi\, \rho(\xi, t) . \tag{26}$$

Then

$$\rho_i(t) = \rho_i(0) + \sum_{j(j\neq i)} P_{ji}\, \rho_j(0) - \rho_i(0) \sum_{j(j\neq i)} P_{ij} = \sum_{j=1}^{\infty} P_{ji}\, \rho_j(0) . \tag{27}$$

It may be assumed without loss of generality that the regions are numbered so that

$$\rho_1(0) \geq \rho_2(0) \geq \rho_3(0) \geq \ldots . \tag{28}$$

Writing $y_i = \rho_i(0)$, $x_i = F(\rho_i(0))$, $\alpha_{ij} = P_{ji}$, and $y_i' = \rho_i(t)$, the premises (3)–(8) of the inequality (9), together with the assumptions about convergence, are recovered and so, provided F is such that both sides converge,

$$V\sum_{i=1}^{\infty} \rho_i(t)F(\rho_i(0)) \geq V\sum_{i=1}^{\infty} \rho_i(0)F(\rho_i(0)) . \tag{29}$$

The inequality (29) has been derived under the assumption that the initial distribution function $\rho(\xi, 0)$ is uniform within each region. An arbitrary $\rho(\xi, 0)$, however, can be approximated by replacing it in each region by its mean value in that region. In this case the right-hand side of (29) is an approximation to the integral $\int d\xi\, \rho(\xi, t)F(\rho(\xi, 0))$, i.e., to the right-hand side of (21); and the left-hand side of (29) is an approximation to the left-hand side of (21). Since the error incurred by these approximations tends to zero for a suitable sequence of subdivisions in which the volume V tends to zero, the inequality (21) must hold.

The discussion of Secs. III and IV remains valid for classical mechanics provided Tr is replaced by integration over the whole of phase space, and the finite-dimensional case of Sec. III is omitted. The remarks in Sec. V remain valid, except that references to unitarity should be replaced by references to Liouville's theorem.

ACKNOWLEDGMENTS

I wish to thank Graham Derrick for many helpful conversations on this subject, Sandy Donald and Sture Nordholm for useful comments on the paper, and Professor H. Messel and the Science Foundation for Physics within the University of Sydney for general support.

APPENDIX A

To prove the inequality (9) from (3)–(8) and the convergence assumptions, we first prove

$$y_1' + y_2' + \cdots + y_k' \leq y_1 + y_2 + \cdots + y_k \tag{A1}$$

for any k, adapting a method due to Ostrowski.[6]

$$\sum_{i=1}^{k} y_i' \equiv \sum_{i=1}^{k}\left(\sum_{j=1}^{\infty} \alpha_{ij} y_j\right) = \sum_{j=1}^{\infty}\left(\sum_{i=1}^{k} \alpha_{ij}\right) y_j . \tag{A2}$$

The change in the order of summation is allowed since, for the matrix whose i,j element is $\alpha_{ij}y_j$, the row sums (sums over j) converge absolutely, and the sum of the row sums, being a finite sum, also converges absolutely.[7] Defining

$$t_j^{(k)} = \sum_{i=1}^{k} \alpha_{ij} \text{ from (A2)},$$

$$\sum_{i=1}^{k} y_i' = \sum_{j=1}^{\infty} t_j^{(k)} y_j . \tag{A3}$$

Further,

$$\sum_{j=1}^{\infty} t_j^{(k)} = \sum_{j=1}^{\infty}\sum_{i=1}^{k} \alpha_{ij} = \sum_{i=1}^{k}\sum_{j=1}^{\infty} \alpha_{ij} = k , \tag{A4}$$

and

$$0 \leq t_j^{(k)} \leq 1 . \tag{A5}$$

Hence

$$\sum_{i=1}^{k} y_i' - \sum_{i=1}^{k} y_i = \sum_{m=1}^{k-1} (t_m^{(k)} - 1)(y_m - y_k) + \sum_{m=k+1}^{\infty} t_m^{(k)}(y_m - y_k) \leq 0 , \tag{A6}$$

which establishes (A1).

Now let us define

$$\xi_1^{(k)} = x_1 - x_2, \quad \xi_2^{(k)} = x_2 - x_3, \ldots$$

$$\xi_k^{(k)} = x_k - x_{k+1} . \tag{A7}$$

Then

$$\zeta_j^{(k)} \leq 0, \quad j=1,2,\ldots,k. \tag{A8}$$

Hence, from (A1),

$$\zeta_1^{(k)}y_1' + \zeta_2^{(k)}(y_1'+y_2') + \cdots + \zeta_k^{(k)}(y_1'+y_2'+\cdots+y_k')$$
$$\geq \zeta_1^{(k)}y_1 + \zeta_2^{(k)}(y_1+y_2) + \cdots + \zeta_k^{(k)}(y_1+y_2+\cdots+y_k),$$

i.e.,

$$y_1'x_1 + y_2'x_2 + \cdots + y_k'x_k \geq y_1x_1 + y_2x_2 + \cdots + y_kx_k + x_{k+1}\left(\sum_{i=1}^k y_i' - \sum_{i=1}^k y_i\right). \tag{A9}$$

If $x_n \leq 0$ for all n the last term is non-negative and

$$\sum_{i=1}^k y_i'x_i \geq \sum_{i=1}^k y_ix_i$$

for all k, proving (9) in this case.

If $x_{n_0} > 0$ for some n_0, then by (3), $x_n \geq x_{n_0} > 0$ for $n \geq n_0$. By hypothesis, $\sum_i y_i$ converges, and it follows from (A1) and the non-negative property of the y_i' that $\sum_i y_i'$ converges absolutely. $\sum_i \alpha_{ij}y_j$ is also absolutely convergent, and so the change in order of summation in the following equation is allowed:

$$\sum_{i=1}^\infty y_i' = \sum_{i=1}^\infty \sum_{j=1}^\infty \alpha_{ij}y_j = \sum_{j=1}^\infty \sum_{i=1}^\infty \alpha_{ij}y_j = \sum_{j=1}^\infty y_j. \tag{A10}$$

Hence the last term on the right-hand side of (A9)

$$\Delta^{(k)} \equiv x_{k+1}\left(\sum_{i=1}^k y_i' - \sum_{i=1}^k y_i\right) = x_{k+1}\left(\sum_{i=k+1}^\infty y_i - \sum_{i=k+1}^\infty y_i'\right). \tag{A11}$$

For $k+1 \geq n_0$,

$$\left| x_{k+1} \sum_{i=k+1}^\infty y_i \right| \leq \left| \sum_{i=k+1}^\infty x_iy_i \right|, \tag{A12}$$

and since $\sum_i x_iy_i$ converges, the right-hand side of (A12), and therefore the left-hand side of (A12), tends to zero as $k \to \infty$. Similarly, $|x_{k+1}\sum_{i=k+1}^\infty y_i'| \to 0$ as $k \to \infty$. Hence $\Delta^{(k)} \to 0$ as $k \to \infty$. This completes the proof of (9).

APPENDIX B

If the density matrix of a system is a monotonic nonincreasing function f of the Hamiltonian H, $\rho = f(H)$, and if the system can also be divided into two independent parts, both of which have an energy spectrum wholly continuous and unbounded above, then the density matrix of the system as a whole is canonical, $f(x) = \exp(-\beta x)$, and the density matrix of each part is canonical with the same β.

Let the density matrices of the two parts be ρ_1 and ρ_2, respectively, the energies E_1 and E_2. Then the monotonic condition for the system as a whole asserts

$$\rho_1(E_1)\rho_2(E_2) \geq \rho_1(E_1')\rho_2(E_2') \quad \text{if } E_1+E_2 \leq E_1'+E_2'. \tag{B1}$$

Writing $E_1 = x$, $E_2 = y$, $F = -\ln\rho_1$, $G = -\ln\rho_2$, Eq. (B1) may be written

$$F(x)+G(y) \leq F(x')+G(y') \quad \text{if } x+y \leq x'+y'. \tag{B2}$$

Let x_0 and $x_0 - \zeta < x_0$ belong to the spectrum of x, and let y_0 belong to the spectrum of y. Then $y_0+\zeta$ belongs to the spectrum of y. By taking, in (B2), first $x=x_0$, $y=y_0$, $x'=x_0-\zeta$, $y'=y_0+\zeta$ and then $x=x_0-\zeta$, $y=y_0+\zeta$, $x'=x_0$, $y'=y_0$ we obtain

$$F(x_0) - F(x_0-\zeta) = G(y_0+\zeta) - G(y_0). \tag{B3}$$

Equation (B3) implies that the derivatives of $F(x)$ and $G(y)$ have the same constant value. It follows that ρ_1, ρ_2, and therefore $\rho_1\rho_2$, all have the canonical form with the same value of β.

---

[1]See, for example, G. V. Chester, Rep. Prog. Phys. 26, 411 (1963).

[2]M. Planck, *Treatise on Thermodynamics* (Dover, New York, 1959).

[3]G. H. Wannier, *Statistical Physics* (Wiley, New York, 1966), p. 119.

[4]J. M. Blatt, Prog. Theor. Phys. 22, 745 (1959).

[5]See, for example, Ref. 3.

[6]A. Ostrowski, J. Math. Pures Appl., 31, 257 (1952).

[7]Konrad Knopp, *Theory and Application of Infinite Series*, 2nd English ed. (Blackie, London, 1951), p. 143.