Statistical Physics

G. Falkovich
http://www.weizmann.ac.il/home/fnfal/papers/statphys15short.pdf

More is different (Anderson)

Contents

This is the first part of a graduate one-semester course. It briefly remind what is supposed to be known from the undergraduate courses about thermodynamics and elementary statistical physics, using a bit more sophisticated language. Small-print parts can be skipped at first reading.

1 Thermodynamics (brief reminder)

Physics is an experimental science, and laws appear usually by induction: from particular cases to a general law and from processes to state functions. The latter step requires integration (to pass, for instance, from Newton equation of mechanics to Hamiltonian or from thermodynamic equations of state to thermodynamic potentials). Generally, it is much easier to differentiate then to integrate and so deduction (or postulation approach) is usually much more simple and elegant. It also provides a good vantage point for further applications and generalizations. In such an approach, one starts from postulating some function of the state of the system and deducing from it the laws that govern changes when one passes from state to state. Here such a deduction is presented for thermodynamics following the book H. B. Callen, *Thermodynamics* (John Wiley & Sons, NYC 1965).

1.1 Basic notions

We use macroscopic description so that some degrees of freedom remain hidden. In mechanics, electricity and magnetism we dealt with the explicitly known macroscopic degrees of freedom but in thermodynamics we deal with macroscopic manifestations of the hidden degrees of freedom. When detailed knowledge is unavailable, physicists use symmetries or conservation laws. Thermodynamics studies restrictions on the possible properties of macroscopic matter that follow from the symmetries of the fundamental laws. Therefore, thermodynamics does not predict numerical values but rather sets inequalities and establishes relations among different properties.

The basic symmetry is invariance with respect to time shifts which gives energy conservation[1]. That allows one to introduce the internal energy E.

---

[1] Be careful trying to build thermodynamic description for biological or social-economic systems, since generally they are not time-invariant. For instance, living beings age and the amount of money is not always conserved.

Energy change generally consists of two parts: the energy change of macroscopic degrees of freedom (which we shall call work) and the energy change of hidden degrees of freedom (which we shall call heat). To be able to measure energy changes in principle, we need adiabatic processes where there is no heat exchange. We wish to establish the energy of a given system in states independent of the way they are prepared. We call such states equilibrium, they are those that can be completely characterized by the *static* values of extensive parameters like energy E, volume V and mole number N (number of particles divided by the Avogadro number 6.02×10^{23}). Other extensive quantities may include numbers of different sorts of particles, electric and magnetic moments etc i.e. everything which value for a composite system is a direct sum of the values for the components. For a given system, any two equilibrium states A and B can be related by an adiabatic process either $A \to B$ or $B \to A$, which allows to measure the difference in the internal energy by the work W done by the system. Now, if we encounter a process where the energy change is not equal to minus the work done by the system, we call the difference the heat flux into the system:

$$dE = \delta Q - \delta W \ . \tag{1}$$

This statement is known as the first law of thermodynamics. The energy is a function of state so we use differential, but we use δ for heat and work, which aren't differentials of any function as they refer to particular forms of energy transfer (not energy content).

The basic problem of thermodynamics is the determination of the equilibrium state that eventually results after all internal constraints are removed in a closed composite system. The problem is solved with the help of extremum principle: there exists an extensive quantity S called entropy which is a function of the extensive parameters of any composite system. The values assumed by the extensive parameters in the absence of an internal constraint maximize the entropy over the manifold of constrained equilibrium states. Since the entropy is extensive it is a homogeneous first-order function of the extensive parameters: $S(\lambda E, \lambda V, \ldots) = \lambda S(E, V, \ldots)$. The entropy is a continuous differentiable function of its variables. This function (called also fundamental relation) is *everything* one needs to know to solve the basic problem (and other problems in thermodynamics as well).

Since the entropy is generally a monotonic function of energy[2] then $S =$

---

[2] This is not always so, we shall see in the second Chapter that the two-level system

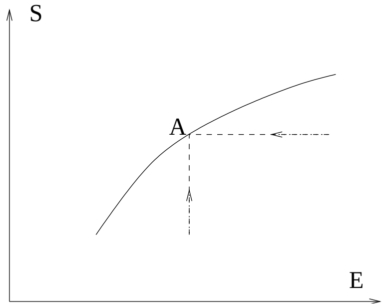
$S(E,V,\ldots)$ can be solved uniquely for $E(S,V,\ldots)$ which is an equivalent fundamental relation. Indeed, assume $(\partial E/\partial S)_X > 0$ and consider $S(E,X)$ and $E(S,X)$. Then[3]

$$\left(\frac{\partial S}{\partial X}\right)_E = 0 \Rightarrow \left(\frac{\partial E}{\partial X}\right)_S = -\frac{\partial(ES)}{\partial(XS)}\frac{\partial(EX)}{\partial(EX)} = -\left(\frac{\partial S}{\partial X}\right)_E \left(\frac{\partial E}{\partial S}\right)_X = 0\ .$$

Differentiating the last relation one more time we get

$$(\partial^2 E/\partial X^2)_S = -(\partial^2 S/\partial X^2)_E(\partial E/\partial S)_X\ ,$$

since the derivative of the second factor is zero as it is at constant X. We thus see that the equilibrium is defined by the energy minimum instead of the entropy maximum (very much like circle can be defined as the figure of either maximal area for a given perimeter or of minimal perimeter for a given area). On the figure, unconstrained equilibrium states lie on the curve while all other states lie below. One can reach the state A either maximizing entropy at a given energy or minimizing energy at a given entropy:



One can work either in energy or entropy representation but ought to be careful not to mix the two.

Experimentally, one usually measures *changes* thus finding derivatives (called equations of state). The partial derivatives of an extensive variable with respect to its arguments (also extensive parameters) are intensive parameters[4]. For example, for the energy one writes

$$\frac{\partial E}{\partial S} \equiv T(S,V,N)\ , \quad \frac{\partial E}{\partial V} \equiv -P(S,V,N) \quad \frac{\partial E}{\partial N} \equiv \mu(S,V,N)\ ,\ldots \tag{2}$$

---

gives a counter-example as well as other systems with a finite phase space.

[3] An efficient way to treat partial derivatives is to use jacobians $\partial(u,v)/\partial(x,y) = (\partial u/\partial x)(\partial v/\partial y) - (\partial v/\partial x)(\partial u/\partial y)$ and the identity $(\partial u/\partial x)_y = \partial(u,y)/\partial(x,y)$.

[4] In thermodynamics we have only extensive and intensive variables (and not, say, surface-dependent terms $\propto N^{2/3}$) because we take thermodynamic limit $N \to \infty$, $V \to \infty$ keeping N/V finite.

These relations are called the *equations of state* and they serve as *definitions* for temperature T, pressure P and chemical potential μ while the respective extensive variables are S, V, N. From (2) we write

$$dE = \delta Q - \delta W = TdS - PdV + \mu dN \ . \tag{3}$$

Entropy is thus responsible for hidden degrees of freedom (i.e. heat) while other extensive parameters describe macroscopic degrees of freedom. The derivatives (2) are defined only in equilibrium. Therefore, $\delta Q = TdS$ and $\delta W = PdV - \mu dN$ for *quasi-static processes* i.e such that the system is close to equilibrium at every point of the process. A process can be considered quasi-static if its typical time of change is larger than the relaxation times (which for pressure can be estimates as L/c, for temperature as L^2/κ, where L is a system size, c - sound velocity and κ thermal conductivity). Finite deviations from equilibrium make $dS > \delta Q/T$ because entropy can increase without heat transfer.

Let us give an example how the entropy maximum principle solves the basic problem. Consider two simple systems separated by a rigid wall which is impermeable for anything but heat. The whole composite system is closed that is $E_1 + E_2 =$const. The entropy change under the energy exchange,

$$dS = \frac{\partial S_1}{\partial E_1} dE_1 + \frac{\partial S_2}{\partial E_2} dE_2 = \frac{dE_1}{T_1} + \frac{dE_2}{T_2} = \left(\frac{1}{T_1} - \frac{1}{T_2}\right) dE_1 \, ,$$

must be positive which means that energy flows from the hot subsystem to the cold one ($T_1 > T_2 \Rightarrow \Delta E_1 < 0$). We see that our definition (2) is in agreement with our intuitive notion of temperature. When equilibrium is reached, $dS = 0$ which requires $T_1 = T_2$. If fundamental relation is known, then so is the function $T(E, V)$. Two equations, $T(E_1, V_1) = T(E_2, V_2)$ and $E_1 + E_2 =$const completely determine E_1 and E_2. In the same way one can consider movable wall and get $P_1 = P_2$ in equilibrium. If the wall allows for particle penetration we get $\mu_1 = \mu_2$ in equilibrium.

Both energy and entropy are homogeneous first-order functions of its variables: $S(\lambda E, \lambda V, \lambda N) = \lambda S(E, V, N)$ and $E(\lambda S, \lambda V, \lambda N) = \lambda E(S, V, N)$ (here V and N stand for the whole set of extensive macroscopic parameters). Differentiating the second identity with respect to λ and taking it at $\lambda = 1$ one gets the Euler equation

$$E = TS - PV + \mu N \ . \tag{4}$$

Let us show that there are only two independent parameters for a simple one-component system, so that chemical potential μ, for instance, can be found as a function of T and P. Indeed, differentiating (4) and comparing with (3) one gets the so-called Gibbs-Duhem relation (in the energy representation) $Nd\mu = -SdT + VdP$ or for quantities per mole, $s = S/N$ and $v = V/N$: $d\mu = -sdT + vdP$. In other words, one can choose $\lambda = 1/N$ and use first-order homogeneity to get rid of N variable, for instance, $E(S,V,N) = NE(s,v,1) = Ne(s,v)$. In the entropy representation,

$$S = E\frac{1}{T} + V\frac{P}{T} - N\frac{\mu}{T} ,$$

the Gibbs-Duhem relation is again states that because $dS = (dE + PdV - \mu dN)/T$ then the sum of products of the extensive parameters and the differentials of the corresponding intensive parameters vanish:

$$Ed(1/T) + Vd(P/T) - Nd(\mu/T) = 0 \ . \tag{5}$$

One uses $\mu(P,T)$, for instance, when considering systems in the external field. One then adds the potential energy (per particle) $u(\mathbf{r})$ to the chemical potential so that the equilibrium condition is $\mu(P,T) + u(\mathbf{r})$ =const. Particularly, in the gravity field $u(\mathbf{r}) = mgz$ and differentiating $\mu(P,T)$ under $T = const$ one gets $vdP = -mgdz$. Introducing density $\rho = m/v$ one gets the well-known hydrostatic formula $P = P_0 - \rho gz$. For composite systems, the number of independent intensive parameters (thermodynamic degrees of freedom) is the number of components plus one. For example, for a mixture of gases, we need to specify the concentration of every gas plus temperature, which is common for all.

Processes. While thermodynamics is fundamentally about states it is also used for describing processes that connect states. Particularly important questions concern performance of engines and heaters/coolers. Heat engine works by delivering heat from a reservoir with some higher T_1 via some system to another reservoir with T_2 doing some work in the process[5]. If the entropy of the hot reservoir decreases by some ΔS_1 then the entropy of the cold one must increase by some $\Delta S_2 \geq \Delta S_1$. The work ΔW is the difference between the heat given by the hot reservoir $\Delta Q_1 = T_1 \Delta S_1$ and the heat absorbed by

---

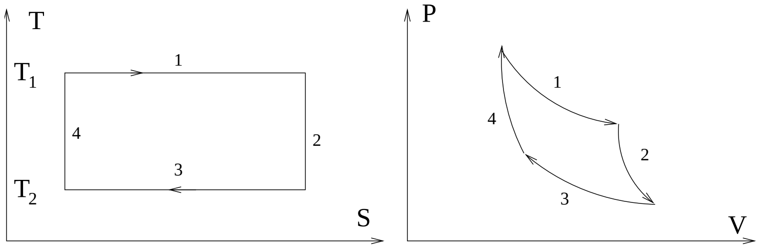
[5] Think about how any real internal combustion engine works to appreciate the level of idealization achieved in distillation of that definition

the cold one $\Delta Q_2 = T_2 \Delta S_2$ (assuming both processes quasi-static). Engine efficiency is the fraction of heat used for work that is

$$\frac{\Delta W}{\Delta Q_1} = \frac{\Delta Q_1 - \Delta Q_2}{\Delta Q_1} = 1 - \frac{T_2 \Delta S_2}{T_1 \Delta S_1} \le 1 - \frac{T_2}{T_1} \ .$$

It is clear that maximal work is achieved for minimal entropy change $\Delta S_2 = \Delta S_1$, which happens for reversible (quasi-static) processes — if, for instance, the system is a gas which works by moving a piston then the pressure of the gas and the work are less for a fast-moving piston than in equilibrium. Similarly, refrigerator/heater is something that does work to transfer heat from cold to hot systems. The performance is characterized by the ratio of transferred heat to the work done. For the cooler, the efficiency is $\Delta Q_2/\Delta W \le T_2/(T_1 - T_2)$, for the heater it is $\Delta Q_1/\Delta W \le T_1/(T_1 - T_2)$. When the temperatures are close, the efficiency is large, as it requires almost no work to transfer heat.

A specific procedure to accomplish reversible heat and work transfer is to use an auxiliary system which undergoes so-called Carnot cycle, where heat exchanges take place only at two temperatures. Engine goes through: 1) isothermal expansion at T_1, 2) adiabatic expansion until temperature falls to T_2, 3) isothermal compression until the entropy returns to its initial value, 4) adiabatic compression until the temperature reaches T_1. The auxiliary system is connected to the reservoirs during isothermal stages: to the first reservoir during 1 and to the second reservoir during 3. During all the time it is connected to our system on which it does work during 1 and 2, increasing the energy of our system, which then decreases its energy by working on the auxiliary system during 3 and 4. The total work is the area of the rectangle between the lines 1,3, the heat ΔQ_1 is the area below the line 1. For heat transfer, one reverses the direction.



Carnot cycle in T-S and P-V variables

Carnot cycle provides one with an operational method to measure the ratio of two temperatures by measuring the engine efficiency[6].

Summary of formal structure. The fundamental relation (in energy representation) $E = E(S, V, N)$ is equivalent to the three equations of state (2). If only two equations of state are given then Gibbs-Duhem relation may be integrated to obtain the third up to an integration constant; alternatively one may integrate molar relation $de = Tds - Pdv$ to get $e(s, v)$ again with an undetermined constant of integration.

Example: consider an ideal monatomic gas characterized by two equations of state (found, say, experimentally with $R \simeq 8.3\,\mathrm{J/mole\,K} \simeq 2\,\mathrm{cal/mole\,K}$):

$$PV = NRT \ , \qquad E = 3NRT/2 \ . \tag{6}$$

The extensive parameters here are E, V, N so we want to find the fundamental equation in the entropy representation, $S(E, V, N)$. We write (4) in the form

$$S = E\frac{1}{T} + V\frac{P}{T} - N\frac{\mu}{T} \ . \tag{7}$$

Here we need to express intensive variables $1/T, P/T, \mu/T$ via extensive variables. The equations of state (6) give us two of them:

$$\frac{P}{T} = \frac{NR}{V} = \frac{R}{v} \ , \qquad \frac{1}{T} = \frac{3NR}{2E} = \frac{3R}{e} \ . \tag{8}$$

Now we need to find μ/T as a function of e, v using Gibbs-Duhem relation in the entropy representation (5). Using the expression of intensive via extensive variables in the equations of state (8), we compute $d(1/T) = -3Rde/2e^2$ and $d(P/T) = -Rdv/v^2$, and substitute into (5):

$$\begin{aligned} d\left(\frac{\mu}{T}\right) &= -\frac{3}{2}\frac{R}{e}de - \frac{R}{v}dv \, , \quad \frac{\mu}{T} = C - \frac{3R}{2}\ln e - R\ln v \, , \\ s &= \frac{1}{T}e + \frac{P}{T}v - \frac{\mu}{T} = s_0 + \frac{3R}{2}\ln\frac{e}{e_0} + R\ln\frac{v}{v_0} \ . \end{aligned} \tag{9}$$

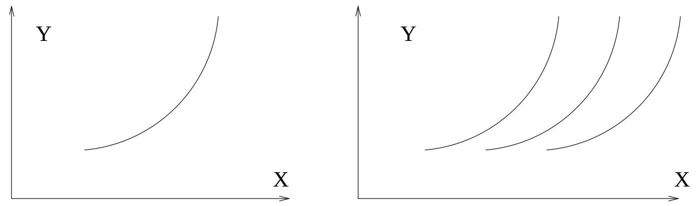
Here e_0, v_0 are parameters of the state of zero internal energy used to determine the temperature units, and s_0 is the constant of integration.

---

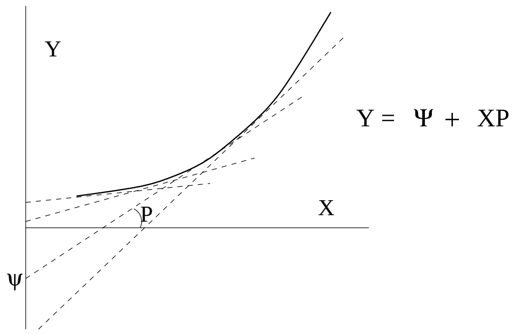
[6] Practical needs to estimate the engine efficiency during the industrial revolution led to the development of such abstract concepts as entropy

1.2 Legendre transform

Let us emphasize that the fundamental relation always relates extensive quantities. Therefore, even though it is always possible to eliminate, say, S from $E = E(S, V, N)$ and $T = T(S, V, N)$ getting $E = E(T, V, N)$, this *is not* a fundamental relation and it does not contain all the information. Indeed, $E = E(T, V, N)$ is actually a partial differential equation (because $T = \partial E/\partial S$) and even if it can be integrated the result would contain undetermined function. Still, it is easier to measure, say, temperature than entropy so it is convenient to have a complete formalism with intensive parameters as operationally independent variables and extensive parameters as derived quantities. This is achieved by the Legendre transform: To pass from the relation $Y = Y(X)$ to that in terms of $P = \partial Y/\partial X$ it is not enough to eliminate X and consider the function $Y = Y(P)$, which determines the curve $Y = Y(X)$ only up to a shift along X:



For example, the same $Y = P^2/4$ correspond to the family of functions $Y = (X + C)^2$ for arbitrary C. To fix the shift one may consider the curve as the envelope of the family of the tangent lines characterized by the slope P and the position ψ of intercept of the Y-axis. The function $\psi(P) = Y[X(P)] - PX(P)$ completely defines the curve; here one substitutes $X(P)$ found from $P = \partial Y(X)/\partial X$ (which is possible only when $\partial P/\partial X = \partial^2 Y/\partial X^2 \neq 0$). The function $\psi(P)$ is referred to as a Legendre transform of $Y(X)$. From $d\psi = -PdX - XdP + dY = -XdP$ one gets $-X = \partial\psi/\partial P$ i.e. the inverse transform is the same up to a sign: $Y = \psi + XP$. In mechanics, we use the Legendre transform to pass from Lagrangian to Hamiltonian description.



Different thermodynamics potentials suitable for different physical situations are obtained replacing different extensive parameters by the respective intensive parameters.

Free energy $F = E - TS$ (also called Helmholtz potential) is that partial Legendre transform of E which replaces the entropy by the temperature as an independent variable: $dF(T, V, N, \ldots) = -SdT - PdV + \mu dN + \ldots$. It is particularly convenient for the description of a system in a thermal contact with a heat reservoir because then the temperature is fixed and we have one variable less to care about. The maximal work that can be done under a constant temperature (equal to that of the reservoir) is minus the differential of the free energy. Indeed, this is the work done *by the system and the thermal reservoir.* That work is equal to the change of the total energy

$$d(E + E_r) = dE + T_r dS_r = dE - T_r dS = d(E - T_r S) = d(E - TS) = dF \ .$$

In other words, the free energy $F = E - TS$ is that part of the internal energy which is *free* to turn into work, the rest of the energy TS we must keep to sustain a constant temperature. The equilibrium state minimizes F, not absolutely, but over the manifold of states with the temperature equal to that of the reservoir. Indeed, consider $F(T, X) = E[S(T, X), X] - TS(T, X)$, then $(\partial E/\partial X)_S = (\partial F/\partial X)_T$ that is they turn into zero simultaneously. Also, in the point of extremum, one gets $(\partial^2 E/\partial X^2)_S = (\partial^2 F/\partial X^2)_T$ i.e. both E and F are minimal in equilibrium. Monatomic gas at fixed T, N has $F(V) = E - TS(V) = -NRT \ln V + \text{const}$. If a piston separates equal amounts then the work done in changing the volume of a subsystem from V_1 to V_2 is $\Delta F = NRT \ln[V_2(V - V_2)/V_1(V - V_1)]$.

Enthalpy $H = E + PV$ is that partial Legendre transform of E which replaces the volume by the pressure $dH(S, P, N, \ldots) = TdS + VdP + \mu dN + \ldots$. It is particularly convenient for situation in which the pressure is maintained constant by a pressure reservoir (say, when the vessel is open into atmosphere). Just as the energy acts as a potential at constant entropy and the free energy as potential at constant temperature, so the enthalpy is a potential for the work done *by the system and the pressure reservoir* at constant pressure. Indeed, now the reservoir delivers pressure which can change the volume so that the differential of the total energy is

$$d(E+E_r)=dE-P_rdV_r=dE+P_rdV=d(E+P_rV)=d(E+PV)=dH\ .$$

Equilibrium minimizes H under the constant pressure. On the other hand, the heat received by the system at constant pressure (and N) is the enthalpy change: $\delta Q = dQ = TdS = dH$. Compare it with the fact that the heat received by the system at constant volume (and N) is the energy change since the work is zero.

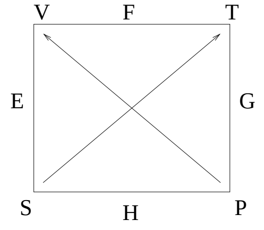
One can replace both entropy and volume obtaining (Gibbs) thermodynamics potential $G = E - TS + PV$ which has $dG(T,P,N,\ldots) = -SdT + VdP + \mu dN + \ldots$ and is minimal in equilibrium at constant temperature and pressure. From (4) we get (remember, they all are functions of different variables):

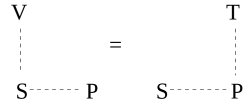
$$F=-P(T,V)V+\mu(T,V)N\ ,\quad H=TS+\mu N\ ,\quad G=\mu(T,P)N\ . \tag{10}$$

When there is a possibility of change in the number of particles (because our system is in contact with some particle source having a fixed chemical potential) then it is convenient to use the grand canonical potential $\Omega(T,V,\mu) = E - TS - \mu N$ which has $d\Omega = -SdT - PdV - Nd\mu$. The grand canonical potential reaches its minimum under the constant temperature and chemical potential.

Since the Legendre transform is invertible, all potentials are equivalent and contain the same information. The choice of the potential for a given physical situation is that of convenience: we usually take what is fixed as a variable to diminish the number of effective variables.

Maxwell relations. Changing order of taking mixed second derivatives of a potential creates a class of identities known as Maxwell relations. For example, $\partial^2E/\partial S\partial V = \partial^2E/\partial V\partial S$ gives $(\partial P/\partial S)_V = -(\partial T/\partial V)_S$. That can be done for all three combinations (SV, SN, VN) possible for a simple single-component system and also for every other potential (F, H, G). Maxwell relations for constant N can be remembered with the help of the mnemonic diagram with the sides labelled by the four common potentials flanked by their respective natural independent variables. In the differential expression for each potential in terms of the natural variables arrow pointing away from the variable implies a positive sign while pointing towards the variable implies negative sign like in $dE = TdS - PdV$:





Maxwell relations are given by the corners of the diagram, for example, $(\partial V/\partial S)_P = (\partial T/\partial P)_S$ etc. If we consider constant N then any fundamental relation of a single-component system is a function of only two variables and therefore have only three independent second derivatives. Traditionally, all derivatives are expressed via the three basic ones (those of Gibbs potential), the specific heat and the coefficient of thermal expansion, both at a constant pressure, and isothermal compressibility:

$$c_P = T\left(\frac{\partial S}{\partial T}\right)_P = -T\left(\frac{\partial^2 G}{\partial T^2}\right)_P , \ \alpha = \frac{1}{V}\left(\frac{\partial V}{\partial T}\right)_P , \ \kappa_T = -\frac{1}{V}\left(\frac{\partial V}{\partial P}\right)_T .$$

In particular, the specific heat at constant volume is as follows:

$$c_V = T\left(\frac{\partial S}{\partial T}\right)_V = c_P - \frac{TV\alpha^2}{N\kappa_T} . \tag{11}$$

That and similar formulas form a technical core of thermodynamics and the art of deriving them ought to be mastered. It involves few simple rules in treating partial derivatives:

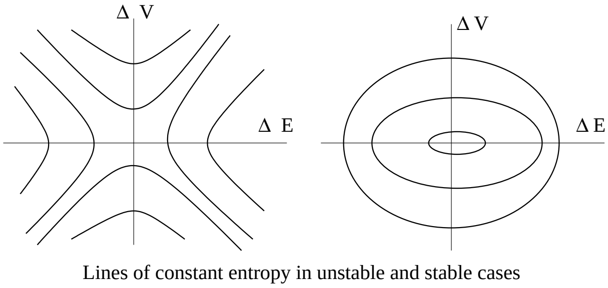
$$\left(\frac{\partial X}{\partial Y}\right)_Z = \left(\frac{\partial Y}{\partial X}\right)_Z^{-1} , \ \left(\frac{\partial X}{\partial Y}\right)_Z\left(\frac{\partial Y}{\partial W}\right)_Z = \left(\frac{\partial X}{\partial W}\right)_Z , \ \left(\frac{\partial X}{\partial Y}\right)_Z\left(\frac{\partial Y}{\partial Z}\right)_X\left(\frac{\partial Z}{\partial X}\right)_Y = -1 .$$

An alternative (and more general) way to manipulate thermodynamic derivatives is to use jacobians and identity $\partial(T,S)/\partial(P,V) = 1$. Taking, say, S, V as independent variables,

$$\frac{\partial(T,S)}{\partial(P,V)} = \frac{\partial(T,S)}{\partial(S,V)}\frac{\partial(S,V)}{\partial(P,V)} = -\frac{(\partial T/\partial V)_S}{(\partial P/\partial S)_V} = \frac{E_{SV}}{E_{VS}} = 1 .$$

1.3 Stability of thermodynamic systems

Consider entropy representation. Stationarity of equilibrium requires $dS = 0$ while stability requires $d^2S < 0$. In particular, that means concavity of $S(E, X)$. Indeed, for all ΔE one must have $S(E+\Delta E, X)+S(E-\Delta E, X) \leq 2S(E, X)$ otherwise our system can break into two halves with the energies $E \pm \Delta E$ thus increasing total entropy. For $\Delta E \to 0$ the stability requirement means $(\partial^2 S/\partial E^2)_X \leq 0 \Rightarrow (\partial T/\partial E)_X \geq 0$ — increase of the energy must increase temperature. For the case $X = V$ this can be also recast into $(\partial T/\partial E)_V = [\partial(TV)/\partial(EV)][\partial(SV)/\partial(SV)] = T^{-1}(\partial T/\partial S)_V = 1/c_v \geq 0$ (adding heat to a stable system increases temperature). The same concavity requirement is true with respect to changes in other parameters X, in particular, $(\partial^2 S/\partial V^2)_E \leq 0 \Rightarrow (\partial P/\partial V)_T \leq 0$ that is isothermal expansion must reduce pressure for the stable system. Considering both changes together we must require $S_{EE}(\Delta E)^2 + 2S_{EV}\Delta E\Delta V + S_{VV}(\Delta V)^2 \leq 0$. This quadratic form, $S_{EE}(\Delta E)^2 + 2S_{EV}\Delta E\Delta V + S_{VV}(\Delta V)^2 = S_{EE}^{-1}(S_{EE}\Delta E + S_{EV}\Delta V)^2 + (S_{VV} - S_{EV}^2 S_{EE}^{-1})(\Delta V)^2$, has a definite sign if the determinant is positive: $S_{EE}S_{VV} - S_{EV}^2 \geq 0$. Manipulating derivatives one can show that this is equivalent to $(\partial P/\partial V)_S \leq 0$. Alternatively, one may consider the energy representation, here stability requires the energy minimum which gives $E_{SS} = T/c_v \geq 0$, $E_{VV} = -(\partial P/\partial V)_S \geq 0$. Considering both variations one can diagonalize $d^2E = E_{SS}(dS)^2 + E_{VV}(dV)^2 + 2E_{SV}dSdV$ by introducing the temperature differential $dT = E_{SS}dS + E_{SV}dV$ so that $2d^2E = E_{SS}^{-1}(dT)^2 + (E_{VV} - E_{SV}^2 E_{SS}^{-1})(dV)^2$. It is thus clear that $E_{VV} - E_{SV}^2 E_{SS}^{-1} = (\partial^2 E/\partial V^2)_T = -(\partial P/\partial V)_T$ and we recover all the same inequalities. Note that the pressure must decrease under both isothermal and adiabatic expansion.

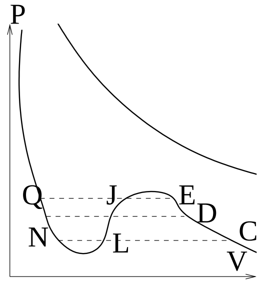


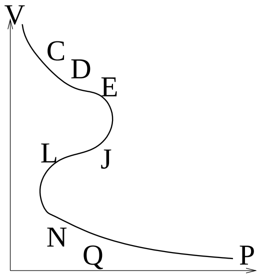
Lines of constant entropy in unstable and stable cases

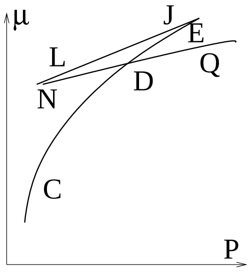
The physical content of those stability criteria is known as Le Chätelier's principle: if some perturbation deviates the system from a stable equilibrium that induces spontaneous processes that reduce the perturbation.

The third law of thermodynamics (Nernst theorem). It claims that $S \to 0$ as $T \to 0$. A standard argument is that since stability requires the positivity of the specific heat c_v then the energy must monotonously increase with the temperature and zero temperature corresponds to the ground state. If the ground state is non-degenerate (unique) then $S = 0$. The ground can be degenerate yet generally that degeneracy grows slower than exponentially with N, then the entropy per particle is zero in the thermodynamic limit. While this argument is correct it is relevant only for temperatures less than the energy difference between the first excited state and the ground state. As such, it has nothing to do with the third law established generally for much higher temperatures and related to the density of states as function of energy. Entropy goes to zero as $T, T^{3/2}, T^3$ for fermions, massive and massless bosons respectively, see Huang (Section 9.4) and L&L (Section 45) for more details.

Phase transitions happen when some stability condition is not satisfied like in the region with $(\partial P/\partial V)_T > 0$ as at the lowest isotherm in the below figure. When the pressure corresponds to the level NLC, it is clear that L is an unstable point and cannot be realized. But which stable point is realized, N or C? To get the answer, one must minimize the Gibbs potential since we have T and P fixed. For one mole, it is the chemical potential which can be found integrating the Gibbs-Duhem relation, $d\mu(T, P) = -sdT + vdP$, under the constant temperature: $G = \mu = \int v(P)dP$. The chemical potential increases up to the point (after E) with infinite dV/dP. After that we move along the isotherm back having $dP < 0$ so that the integral decreases and then passes through another point of infinite derivative and start to increase again. In other words, the third graph below represents three branches of the function $\mu(P)$ that has its derivative the function $v(P)$ shown in the second graph. It is clear that to the intersection point D correspond to equal areas below and above the horizontal line on the first graph.The pressure that corresponds to this point separates the absolute minimum at the left branch marked Q (solid-like) from that on the right one marked C (liquid-like). The dependence of volume on pressure is discontinuous along the isotherm.


P
Q
J
E
D
N
L
C
V


V
C
D
E
L
J
N
Q
P


μ
J
E
L
Q
N
D
C
P

2 Basic statistical physics (brief reminder)

Here we describe two principal ways (microcanonical and canonical) to derive thermodynamics from statistical mechanics and introduce microscopic statistical description in the phase space.

2.1 Microcanonical distribution

Consider a *closed* system with the energy E_0. Boltzmann *assumed* that all microstates with the same energy have equal probability (ergodic hypothesis) which gives the *microcanonical distribution*:

$$\rho(p,q) = A\delta[E(p,q) - E_0] \ . \tag{12}$$

Usually one considers the energy fixed with the accuracy Δ so that the microcanonical distribution is

$$\rho = \begin{cases} 1/\Gamma & \text{for } E \in (E_0, E_0 + \Delta) \\ 0 & \text{for } E \notin (E_0, E_0 + \Delta) \,, \end{cases} \tag{13}$$

where Γ is the volume of the phase space occupied by the system

$$\Gamma(E, V, N, \Delta) = \int_{E<\mathcal{H}<E+\Delta} d^{3N}p d^{3N}q \ . \tag{14}$$

For example, for N noninteracting particles (ideal gas) the states with the energy $E = \sum p^2/2m$ are in the **p**-space near the hyper-sphere with the radius $\sqrt{2mE}$. Remind that the surface area of the hyper-sphere with the radius R in $3N$-dimensional space is $2\pi^{3N/2}R^{3N-1}/(3N/2-1)!$ and we have

$$\Gamma(E, V, N, \Delta) \propto E^{3N/2-1} V^N \Delta/(3N/2-1)! \approx (E/N)^{3N/2} V^N \Delta \ . \tag{15}$$

To link statistical physics with thermodynamics one must define the fundamental relation i.e. a thermodynamic potential as a function of respective variables. It can be done using either canonical or microcanonical distribution. We start from the latter and introduce the entropy as

$$S(E, V, N) = \ln \Gamma(E, V, N) \ . \tag{16}$$

This is one of the most important formulas in physics[7] (on a par with $F = ma\,, E = mc^2$ and $E = \hbar\omega$).

---

[7] It is inscribed on the Boltzmann's gravestone.

Let's see if such definition of entropy makes sense from the viewpoint of thermodynamics. In Section 2 we see that the entropy derivative defines temperature which must be the same for two equilibrium systems in contact. Noninteracting subsystems are statistically independent. That means that the statistical weight of the composite system is a product - indeed, for every state of one subsystem we have all the states of another. If the weight is a product then the entropy is a sum. For interacting subsystems, this is true only for short-range forces in the thermodynamic limit $N \to \infty$. Consider two subsystems, 1 and 2, that can exchange energy. Assume that the indeterminacy in the energy of any subsystem, Δ, is much less than the total energy E. Then a subsystem can have $E_i = i\Delta$ for any $0 \le i \le N$, and the number of states of the composite system is as follows:

$$\Gamma(E) = \sum_{i=1}^{E/\Delta} \Gamma_1(E_i)\Gamma_2(E - E_i) \ . \tag{17}$$

We denote $\bar{E}_1, \bar{E}_2 = E - \bar{E}_1$ the values that correspond to the maximal term in the sum (17). The derivative of it is proportional to $(\partial\Gamma_1/\partial E_i)\Gamma_2 + (\partial\Gamma_2/\partial E_i)\Gamma_1 = (\Gamma_1\Gamma_2)^{-1}[(\partial S_1/\partial E_1)_{\bar{E}_1} - (\partial S_2/\partial E_2)_{\bar{E}_2}]$. Then the extremum condition is evidently $(\partial S_1/\partial E_1)_{\bar{E}_1} = (\partial S_2/\partial E_2)_{\bar{E}_2}$, that is the extremum corresponds to the thermal equilibrium where the temperatures of the subsystems are equal. The equilibrium is thus where the maximum of probability is. It is obvious that $\Gamma(\bar{E}_1)\Gamma(\bar{E}_2) \le \Gamma(E) \le \Gamma(\bar{E}_1)\Gamma(\bar{E}_2)E/\Delta$. If the system consists of N particles and $N_1, N_2 \to \infty$ then $S(E) = S_1(\bar{E}_1) + S_2(\bar{E}_2) + O(logE)$ where the last term is negligible in the thermodynamic limit. We thus reveal the statistical meaning of uniformity of temperature in thermal equilibrium - maximum of probability. States with unequal temperatures of the subsystems are possible but less probable.

Let us now see how identification with the thermodynamic entropy proceeds for a particular system, for instance, an ideal gas. The problem is that the logarithm of (15) contains non-extensive term $N \ln V$. The resolution of this controversy is that to get thermodynamics we need to treat the particles as indistinguishable. That means not to count as different the states where particles are interchanged, that is not to account for the entropy of mixing different species. Indeed, in thermodynamics, we implicitly assume that mixing different parts of the same gas is a reversible process which is only true if the particles are not distinguished. For identical particles, one needs to divide Γ (15) by the number of transmutations $N!$ which makes the resulting entropy of the ideal gas extensive: $S(E,V,N) = (3N/2)\ln E/N + N\ln eV/N$+const. Note that quantum particles (atoms and molecules) are indeed indistinguishable, which is expressed by a proper symmetrization of the wave function. One can only wonder at the genius of Gibbs who introduced $N!$ long before quantum mechanics (see, L&L 40 or Pathria 1.5 and 6.1). One is tempted to ask: but if we consider a system of distinguishable classical bodies, like a set of billiard balls, should the entropy be dependent on whether we mark them? The answer is yes.

Defining temperature in a usual way, $T^{-1} = \partial S/\partial E = 3N/2E$, we get the correct equation of state $E = 3NT/2$. We express here temperature in the energy units. To pass to Kelvin degrees, one transforms $T \to kT$ and $S \to kS$ where the Boltzmann constant $k = 1.38 \cdot 10^{23}\, J/K$. The value of classical entropy (16) depends on the units. Proper quantitative definition comes from quantum physics with Γ being the number of microstates that correspond to a given value of macroscopic parameters. In the quasi-classical limit the number of states is obtained by dividing the phase space into units with $\Delta p \Delta q = 2\pi\hbar$.

The same definition (entropy as a logarithm of the number of states) is true for any system with a discrete set of states. For example, consider the set of N two-level systems with levels 0 and ϵ. If energy of the set is E then there are $L = E/\epsilon$ upper levels occupied. The statistical weight is determined by the number of ways one can choose L out of N: $\Gamma(N,L) = C_N^L = N!/L!(N-L)!$. We can now define entropy (i.e. find the fundamental relation): $S(E,N) = \ln\Gamma$. Considering $N \gg 1$ and $L \gg 1$ we can use the Stirling formula in the form $d\ln L!/dL = \ln L$ and derive the equation of state (temperature-energy relation),

$$T^{-1} = \partial S/\partial E = \epsilon^{-1}\frac{\partial}{\partial L}\ln\frac{N!}{L!(N-L)!} = \epsilon^{-1}\ln\frac{N-L}{L}\ ,$$

and specific heat $C = dE/dT = N(\epsilon/T)^2 2\cosh^{-2}(\epsilon/T)$. Note that the ratio of the number of particles on the upper level to those on the lower level is $L/(N-L) = \exp(-\epsilon/T)$ (Boltzmann relation).

The derivation of thermodynamic fundamental relation $S(E,\ldots)$ in the microcanonical ensemble is thus via the number of states or phase volume.

2.2 Canonical distribution

Let us now derive the canonical distribution from the microcanonical. Consider a small subsystem or a system in a contact with the thermostat (which can be thought of as consisting of infinitely many copies of our system — this is so-called canonical ensemble, characterized by N, V, T). Here our system can have any energy and the question arises what is the probability $W(E)$. Let us find first the probability of the system to be in a given microstate a with the energy E. Since all the states of the thermostat are equally likely to occur, then the probability should be directly proportional to the statistical weight of the thermostat $\Gamma_0(E_0 - E)$. How this quantity depends on E? One obtains a universal answer in the limit $E \ll E_0$, when one can expand $\Gamma_0(E_0 - E)$. Since this is a very sharp function of E, one expands its logarithm (or the function in the exponent), which is slower: $\Gamma_0(E_0 - E) = \exp[S_0(E_0 - E)] \approx \exp[S_0(E_0) - E/T)]$ and obtain

$$w_a(E) = Z^{-1}\exp(-E/T)\ , \qquad (18)$$

$$Z = \sum_a \exp(-E_a/T)\ . \qquad (19)$$

Note that there is no trace of the thermostat left except for the temperature. The normalization factor $Z(T, V, N)$ is a sum over all states accessible to the system and is called the partition function.

The probability to have a given energy is the probability of the state (18) times the number of states i.e. the statistical weight of the *subsystem*:

$$W(E) = \Gamma(E) w_a(E) = \Gamma(E) Z^{-1}\exp(-E/T)\ . \qquad (20)$$

Here the weight $\Gamma(E)$ grows with E very fast for large N. But as $E \to \infty$ the exponent $\exp(-E/T)$ decays faster than any power. As a result, $W(E)$ is concentrated in a very narrow peak and the energy fluctuations around $\bar{E}$ are very small. For example, for an ideal gas $W(E) \propto E^{3N/2}\exp(-E/T)$. Let us stress again that the Gibbs canonical distribution (18) tells that the probability of a given microstate exponentially decays with the energy of the state while (20) tells that the probability of a given energy has a peak.

An alternative and straightforward way to derive the canonical distribution is to use consistently the Gibbs idea of the canonical ensemble as a virtual set, of which the single member is the system under consideration and the energy of the total set is fixed. The probability to have our chosen system in the state a with the energy E_a is then given by the average number of systems $\bar{n}_a$ in this state divided by the total number of systems N. Consider a set of occupation numbers denoted as $\{n_a\} = (n_0, n_1, n_2 \ldots)$, which means that there are n_0 subsystems in the ground state, n_1 subsystems in the first excited state, etc. Every set of the occupation numbers satisfies obvious conditions

$$\sum_a n_a = N \ , \qquad \sum_a E_a n_a = E = \epsilon N \ . \tag{21}$$

Any given set is realized in $W\{n_a\} = N!/n_0!n_1!n_2! \ldots$ number of ways and the probability to realize the set is proportional to the respective W:

$$\bar{n}_a = \frac{\sum n_a W\{n_a\}}{\sum W\{n_a\}} \ , \tag{22}$$

where summation goes over all the sets that satisfy (21). We assume that in the limit when $N, n_a \to \infty$ the main contribution into (22) is given by the most probable distribution that is maximum of W (we actually look at the maximum of $\ln W$ which is the same yet technically simpler) under the constraints (21). Using the method of Lagrangian multipliers we look for the extremum of $\ln W - \alpha \sum_a n_a - \beta \sum_a E_a n_a$. Using the Stirling formula $\ln n! = n \ln n - n$ we write $\ln W = N \ln N - \sum_a n_a \ln n_a$. We thus need to find the value n_a^* which corresponds to the extremum of $\sum_a n_a \ln n_a - \alpha \sum_a n_a - \beta \sum_a E_a n_a$. Differentiating we obtain: $\ln n_a^* = -\alpha - 1 - \beta E_a$ which gives

$$\frac{n_a^*}{N} = \frac{\exp(-\beta E_a)}{\sum_a \exp(-\beta E_a)} \ . \tag{23}$$

The parameter β is given implicitly by the relation

$$\frac{E}{N} = \epsilon = \frac{\sum_a E_a \exp(-\beta E_a)}{\sum_a \exp(-\beta E_a)} \ . \tag{24}$$

Of course, physically $\epsilon(\beta)$ is usually more relevant than $\beta(\epsilon)$. See Pathria, Sect 3.2.

To get thermodynamics from the Gibbs distribution one needs to define the free energy, $F = E - TS$, because we are under a constant temperature. This is done via the partition function Z (which is of central importance since macroscopic quantities are generally expressed via the derivatives of it):

$$F(T, V, N) = -T \ln Z(T, V, N) \ . \tag{25}$$

Indeed, one differentiates the identity $Z = \exp(-F/T) = \sum_a \exp(-E_a/T)$ with respect to temperature and obtains

$$F = \bar{E} + T\left(\frac{\partial F}{\partial T}\right)_V \ ,$$

equivalent to $F = E - TS$ in thermodynamics.

One can also come to this by defining entropy. Remind that for a closed system we defined $S = \ln\Gamma$ while the probability of state was $w_a = 1/\Gamma$. In other words, the entropy was minus the log of probability. Let us generalize this defining entropy as minus the mean log of probability. For a system in contact with a thermostat that has a Gibbs distribution we have $\ln w_a$ linear in E, so that

$$\begin{aligned} S(\bar{E}) &= - \ \ln w_a(\bar{E}) = -\langle \ln w_a \rangle = -\sum w_a \ln w_a \\ &= \ \sum w_a(E_a/T + \ln Z) = E/T + \ln Z = (E-F)/T \ . \end{aligned} \tag{26}$$

Even though we derived the formula for entropy, $S = -\sum w_a \ln w_a$, for an equilibrium, this definition can be used for any set of probabilities w_a, since it provides a useful measure of our ignorance about the system, as we shall see later.

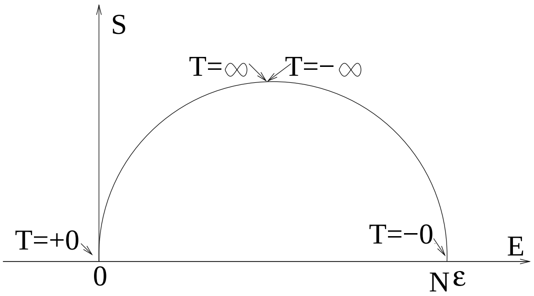
See Landau & Lifshitz (Sects 31,36).

2.3 Two-level system

We have seen that the central element of statistical physics is counting the states. Here we consider an example with the simplest structure of energy levels to illustrate the use of microcanonical and canonical distributions. Assume levels 0 and ϵ. Remind that in Sect. 2.1 we already considered two-level system in the microcanonical approach calculating the number of ways one can distribute $L = E/\epsilon$ portions of energy between N particles and obtaining $S(E,N) = \ln C_N^L = \ln[N!/L!(N-L)!] \approx N\ln[N/(N-L)] + L\ln[(N-L)/L]$. The temperature in the microcanonical approach is as follows:

$$T^{-1} = \frac{\partial S}{\partial E} = \epsilon^{-1}(\partial/\partial L)\ln[N!/L!(N-L)!] = \epsilon^{-1}\ln(N-L)/L \ . \tag{27}$$

The entropy as a function of energy is drawn on the Figure:



Indeed, entropy is zero at $E = 0, N\epsilon$ when all the particles are in the same state. The entropy is symmetric about $E = N\epsilon/2$. We see that when $E > N\epsilon/2$ then the population of the higher level is larger than of the lower one (inverse population as in a laser) and the temperature is negative. Negative temperature may happen only in systems with the upper limit of energy levels and simply means that by adding energy beyond some level we actually decrease the entropy i.e. the number of accessible states. That example with negative temperature is to help you to disengage from the everyday notion of temperature and to get used to the physicist idea of temperature as the derivative of energy with respect to entropy.

Available (non-equilibrium) states lie below the $S(E)$ plot, notice that the entropy maximum corresponds to the energy minimum for positive temperatures and to the energy maximum for the negative temperatures part. A glance on the figure also shows that when the system with a negative temperature is brought into contact with the thermostat (having positive temperature) then our system gives away energy (a laser generates and emits light) decreasing the temperature further until it passes through infinity to positive values and eventually reaches the temperature of the thermostat. That is negative temperatures are actually "hotter" than positive. By itself though the system is stable since $\partial^2 S/\partial E^2 = -N/L(N-L)\epsilon^2 < 0$.

Let us stress that there is no volume in $S(E, N)$ that is we consider only subsystem or only part of the degrees of freedom. Indeed, real particles have kinetic energy unbounded from above and can correspond only to positive temperatures [negative temperature and infinite energy give infinite Gibbs factor $\exp(-E/T)$].

Apart from laser, an example of a two-level system is spin 1/2 in the magnetic field H. Because the interaction between the spins and atom motions (spin-lattice relaxation) is weak then the spin system for a long time (tens of minutes) keeps its separate temperature and can be considered separately.

External fields are parameters (like volume and chemical potential) that determine the energy levels of the system. They are sometimes called generalized thermodynamic coordinates, and the derivatives of the energy with respect to them are called respective forces. Let us derive the generalized force M that corresponds to the magnetic field and determines the work done under the change of magnetic field: $dE(S,H) = TdS - MdH$. Since the projection of every magnetic moment on the direction of the field can take two values $\pm\mu$ then the magnetic energy of the particle is $\mp\mu H$ and $E = -\mu(N_+ - N_-)H$. The force (the partial derivative of the energy with respect to the field at a fixed entropy) is called magnetization or magnetic moment of the system:

$$M = -\left(\frac{\partial E}{\partial H}\right)_S = \mu(N_+ - N_-) = N\mu\frac{\exp(\mu H/T) - \exp(-\mu H/T)}{\exp(\mu H/T) + \exp(-\mu H/T)} . \quad (28)$$

The derivative was taken at constant entropy that is at constant populations N_+ and N_-. Note that negative temperature for the spin system corresponds to the magnetic moment opposite in the direction to the applied magnetic field. Such states are experimentally prepared by a fast reversal of the magnetic field. We can also define magnetic susceptibility: $\chi(T) = (\partial M/\partial H)_{H=0} = N\mu^2/T$, yet another second derivative that determines the response and fluctuations and will feature prominently in what follows.

At weak fields and positive temperature, $\mu H \ll T$, (28) gives the formula for the so-called Pauli paramagnetism

$$\frac{M}{N\mu} = \frac{\mu H}{T} . \quad (29)$$

Para means that the majority of moments point in the direction of the external field. This formula shows in particular a remarkable property of the spin system: adiabatic change of magnetic field (which keeps constant N_+, N_- and thus M) is equivalent to the change of temperature even though spins do not exchange energy. One can say that under the change of the value of the homogeneous magnetic field the relaxation is instantaneous in the spin system. This property is used in cooling the substances that contain paramagnetic impurities. For the entropy of the spin system to be preserved, one needs to change the field slowly comparatively to the spin-spin relaxation and fast comparatively to the spin-lattice relaxation. The first condition means that one cannot reach negative temperatures by adiabatically reversing magnetic field since the relaxation times of spins grow when field decreases; indeed, negative temperatures must be reached through $T \to \infty$, not zero. In practice, negative temperatures were reached (by Purcell, Pound and Ramsey in 1951) by *fast* reversal of the magnetic field.

To conclude let us treat the two-level system by the canonical approach where we calculate the partition function and the free energy:

$$Z(T,N) = \sum_{L=0}^{N} C_N^L \exp[-L\epsilon/T] = [1+\exp(-\epsilon/T)]^N \,, \tag{30}$$

$$F(T,N) = -T\ln Z = -NT\ln[1+\exp(-\epsilon/T)] \,. \tag{31}$$

We can now re-derive the entropy as $S = -\partial F/\partial T$ and derive the (mean) energy and specific heat:

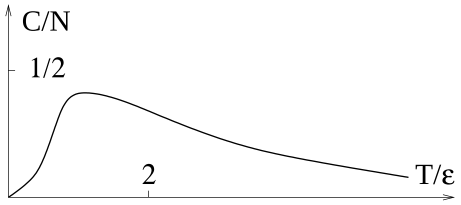
$$\bar{E} = Z^{-1}\sum_a E_a \exp(-\beta E_a) = -\frac{\partial \ln Z}{\partial \beta} = T^2 \frac{\partial \ln Z}{\partial T} \tag{32}$$

$$= \frac{N\epsilon}{1+\exp(\epsilon/T)} \,, \tag{33}$$

$$C = \frac{dE}{dT} = \frac{N\exp(\epsilon/T)}{[1+\exp(\epsilon/T)]^2}\frac{\epsilon^2}{T^2} \,. \tag{34}$$

Here (32) is a general formula which we shall use in the future. Remark that even though canonical approach corresponds to a system in a thermostat, which necessary has positive temperature, all the formulas make sense at negative T too.

Specific heat is one of the second derivatives of the thermodynamic potentials; such quantities characterize the response of the system to change of parameters and will feature prominently in the course. Specific heat tells us how much one raises the energy of the system when increasing the temperature by one degree (or, alternatively, how much energy one needs to increase the temperature by one degree). Specific heat of a two-level system turns into zero both at low temperatures (too small portions of energy are "in circulation") and at high temperatures (occupation numbers of two levels already close to equal so changing temperature does not change energy).



A specific heat of this form characterized by a peak is observed in all systems with an excitation gap.

More details can be found in Kittel, Section 24 and Pathria, Section 3.9.

2.4 Distribution in the phase space

Here we introduce microscopic statistical description in the phase space and re-derive the canonical distribution. The phase space $(\mathbf{p}, \mathbf{q})$ has $6N$ dimensions. Instead of counting states we consider now evolution and define probability for a subsystem to be in some $\Delta\mathbf{p}\Delta\mathbf{q}$ region of the phase space as the fraction of time it spends there: $w = \lim_{T\to\infty} \Delta t/T$. Assuming that the probability to find the subsystem within the volume $d\mathbf{p}d\mathbf{q}$ is proportional to this volume, we introduce the statistical distribution in the phase space as density: $dw = \rho(p,q)d\mathbf{p}d\mathbf{q}$. By definition, the average with the statistical distribution is equivalent to the time average:

$$\bar{f} = \int f(\mathbf{p},\mathbf{q})\rho(\mathbf{p},\mathbf{q})d\mathbf{p}d\mathbf{q} = \lim_{T\to\infty}\frac{1}{T}\int_0^T f(t)dt\ . \tag{35}$$

The main idea is that $\rho(\mathbf{p},\mathbf{q})$ for a subsystem does not depend on the initial states of this and other subsystems so it can be found without actually solving equations of motion. We define statistical equilibrium as a state where macroscopic quantities equal to the mean values. Assuming short-range forces we conclude that different macroscopic subsystems interact weakly and are statistically independent so that the distribution for a composite system ρ_{12} is factorized: $\rho_{12} = \rho_1\rho_2$.

Now, we take the ensemble of identical systems starting from different points in phase space. As coordinates and momenta change with time, each point moves in the phase space. Motion of all points of the ensemble constitutes a flow with the velocity $\mathbf{v} = (\dot{\mathbf{p}}, \dot{\mathbf{q}})$, where the density at every point of the phase space changes according to the continuity equation: $\partial\rho/\partial t + div\,(\rho\mathbf{v}) = 0$. However, more interesting is the density change along the trajectories, i.e. how the probability distribution changes with time for every given system. Such evolution is determined by the time derivative along the flow: $d\rho(t,\mathbf{p},\mathbf{q})/dt = \partial\rho/\partial t + \sum_i(\dot{q}_i\partial\rho/\partial q_i + \dot{p}_i\partial\rho/\partial p_i = \partial\rho/\partial t + (\mathbf{v}\cdot\nabla)\rho$. According to the continuity equation, it is $d\rho(t,\mathbf{p},\mathbf{q})/dt = -\rho div\,\mathbf{v}$. If the motion is considered for not very large time it is conservative and can be described by the Hamiltonian dynamics: $\dot{q}_i = \partial\mathcal{H}/\partial p_i$ and $\dot{p}_i = -\partial\mathcal{H}/\partial q_i$. Here the Hamiltonian generally depends on the momenta and coordinates of the given subsystem and its neighbors. Hamiltonian flow in the phase space is incompressible, it conserves area in each plane p_i, q_i and the total volume: $div\,\mathbf{v} = \partial\dot{q}_i/\partial q_i + \partial\dot{p}_i/\partial p_i = 0$. That gives the Liouville theorem: $d\rho/dt = \partial\rho/\partial t + (\mathbf{v}\cdot\nabla)\rho = -\rho div\,\mathbf{v} = 0$. The statistical distribution is thus conserved along the phase trajectories of any subsystem. As a result, equilibrium distribution ρ is an integral of motion. Consequently, it must be expressed solely via the integrals of motion. Since $\ln\rho$ is an additive quantity then it must be expressed linearly via the additive integrals of motions which for a general mechanical system are energy $E(p,q)$, momentum $\mathbf{P}(p,q)$ and the momentum of momentum $\mathbf{M}(p,q)$:

$$\ln\rho_a = \alpha_a + \beta E_a(p,q) + \mathbf{c}\cdot\mathbf{P}_a(p,q) + \mathbf{d}\cdot\mathbf{M}(p,q)\ . \tag{36}$$

Here α_a is the normalization constant for a given subsystem while the constants $\beta, \mathbf{c}, \mathbf{d}$ are the same for all subsystems to ensure additivity. Those seven constants (in 3 dimensions) are determined by the values of the seven integrals of motion for the whole system. We thus conclude that the additive integrals of motion is all we need to get the statistical distribution of a closed system (and any subsystem), those integrals replace all the enormous microscopic information. Considering system which neither moves nor rotates we are down to the single integral, energy. For any subsystem (or any system in the contact with thermostat) we get Gibbs' *canonical distribution*

$$\rho(p,q) = A\exp[-\beta E(p,q)]\ . \tag{37}$$

See Landau & Lifshitz, Sects 1-4.

3 Entropy and information

By definition, entropy of a closed system determines the number of available states (or, classically, phase volume). Assuming that system spends comparable time in different available states we conclude that since the equilibrium must be the most probable state it corresponds to the entropy maximum. If the system happens to be not in equilibrium at a given moment of time [say, the energy distribution between the subsystems is different from the most probable Gibbs distribution] then it is more probable to go towards equilibrium that is increasing entropy. This is a microscopic (probabilistic) interpretation of the *second law of thermodynamics* formulated by Clausius in 1865. The probability maximum is very sharp in the thermodynamic limit since $\exp(S)$ grows exponentially with the system size. That means that for macroscopic systems the probability to pass into the states with lower entropy is so vanishingly small that such events are never observed.

What often causes confusion here is that the dynamics (classical and quantum) of any given system is time reversible. The Hamiltonian evolution described in Sect 2.4 preserves the density in the phase space $\rho(p, q)$, so how the entropy $S = -\int dpdq\rho \ln \rho$ can grow? To avoid the confusion, one must remember that we study the situations with incomplete knowledge of the system. That means that we know coordinates and momenta within some intervals, i.e. characterize the system not by a point in phase space but by a finite region there. Entropy growth is then related not to the trajectory of a single point in phase space but to the behavior of finite regions (i.e. sets of such points) or ensembles of systems. The entropy of a single point is zero. The necessity to consider finite regions follows from the insufficiency of information about the true state of the system. Consideration of finite regions is called *coarse graining* and it is the main feature of stat-physical approach responsible for the irreversibility of statistical laws. In this section we shall see how it works.

3.1 Lyapunov exponent

The dynamical mechanism of the entropy growth is the separation of trajectories in phase space so that trajectories started from a small finite region are found in larger and larger regions of phase space as time proceeds. The relative motion is determined by the velocity difference between neighboring points in the phase space: $\delta v_i = r_j \partial v_i / \partial x_j = r_j \sigma_{ij}$. Here $\mathbf{x} = (\mathbf{p}, \mathbf{q})$ is the $6N$-dimensional vector of the position and $\mathbf{v} = (\dot{\mathbf{p}}, \dot{\mathbf{q}})$ is the velocity in the phase space. The trace of the tensor σ_{ij} is the rate of the volume change which must be zero according to the Liouville theorem (that is a Hamiltonian dynamics imposes an incompressible flow in the phase space). We can decompose the tensor of velocity derivatives into an antisymmetric part (which describes rotation) and a symmetric part (which describes deformation). We are interested here in deformation because it is the mechanism of the entropy growth. The symmetric tensor, $S_{ij} = (\partial v_i / \partial x_j + \partial v_j / \partial x_i)/2$, can be always transformed into a diagonal form by an orthogonal transformation (i.e. by the rotation of the axes), so that $S_{ij} = S_i \delta_{ij}$. Recall that for Hamiltonian motion, $\sum_i S_i = div\, \mathbf{v} = 0$, so that some components are positive, some are negative. Positive diagonal components are the rates of stretching and negative components are the rates of contraction in respective directions. Indeed, the equation for the distance between two points along a principal direction has a form: $\dot{r}_i = \delta v_i = r_i S_i$. The solution is as follows:

$$r_i(t) = r_i(0) \exp\left[\int_0^t S_i(t')\,dt'\right] . \tag{38}$$

For a time-independent strain, the growth/decay is exponential in time. One recognizes that a purely straining motion converts a spherical element into an ellipsoid with the principal diameters that grow (or decay) in time. Indeed, consider a two-dimensional projection of the initial spherical element i.e. a circle of the radius R at $t = 0$. The point that starts at $x_0, y_0 = \sqrt{R^2 - x_0^2}$ goes into

$$\begin{aligned} & x(t) = e^{S_{11}t} x_0 \,, \\ & y(t) = e^{S_{22}t} y_0 = e^{S_{22}t}\sqrt{R^2 - x_0^2} = e^{S_{22}t}\sqrt{R^2 - x^2(t)e^{-2S_{11}t}} \,, \\ & x^2(t)e^{-2S_{11}t} + y^2(t)e^{-2S_{22}t} = R^2 \,. \end{aligned} \tag{39}$$

The equation (39) describes how the initial circle turns into the ellipse whose eccentricity increases exponentially with the rate $|S_{11} - S_{22}|$. In a multi-dimensional space, any sphere of initial conditions turns into the ellipsoid defined by $\sum_{i=1}^{6N} x_i^2(t) e^{-2S_i t}$ =const.

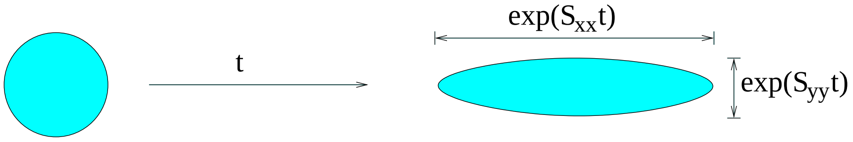


Figure 1: Deformation of a phase-space element by a permanent strain.

Of course, as the system moves in the phase space, both the strain values and the orientation of the principal directions change, so that expanding direction may turn into a contracting one and vice versa. Since we do not want to go into details of how the system interacts with the environment, then we consider such evolution as a kind of random process. The question is whether averaging over all values and orientations gives a zero net result. It may seem counter-intuitive at first, but in a general case an exponential stretching must persist on average and the majority of trajectories separate.

Physicists think in two ways: one in space and another in time (unless they are relativistic and live in a space-time).

Let us first look at separation of trajectories from a temporal perspective: even when the average rate of separation along a given direction $\Lambda_i(t) = \int_0^t S_i(t')dt'/t$ is zero, the average exponent of it is larger than unity (and generally growing with time):

$$\frac{1}{T}\int_0^T dt \exp\left[\int_0^t S_i(t')dt'\right] \geq 1 \ .$$

This is because the intervals of time with positive $\Lambda(t)$ give more contribution into the exponent than the intervals with negative $\Lambda(t)$. That follows from the *concavity* of the exponential function. In the simplest case, when $-a < \Lambda < a$, the average Λ is zero, while the average exponent is $(1/2a)\int_a^{-a} e^\Lambda d\Lambda = (e^a - e^{-a})/2a > 1$.

Looking from a spatial perspective, consider the simplest example of a two-dimensional pure strain, which corresponds to an incompressible saddle-point flow: $v_x = \lambda x$, $v_y = -\lambda y$. Here we have one expanding direction direction and one contracting direction, their rates being equal. The vector $\mathbf{r} = (x, y)$ which characterizes the distance between two close trajectories can look initially at any direction. The evolution of the vector components satisfies the equations $\dot{x} = v_x$ and $\dot{y} = v_y$. Whether the vector is stretched or contracted after some time T depends on its orientation and on T. Since $x(t) = x_0 \exp(\lambda t)$ and $y(t) = y_0 \exp(-\lambda t) = x_0 y_0 / x(t)$ then every trajectory is a hyperbole. A unit vector initially forming an angle φ with the x axis will have its length $[\cos^2\varphi \exp(2\lambda T) + \sin^2\varphi \exp(-2\lambda T)]^{1/2}$ after time T. The vector will be stretched if $\cos\varphi \geq [1 + \exp(2\lambda T)]^{-1/2} < 1/\sqrt{2}$, i.e. the fraction of stretched directions is larger than half. When along the motion all orientations are equally probable, the net effect is stretching, proportional to the persistence time T.

The net stretching and separation of trajectories is formally proved in mathematics by considering random strain matrix $\hat{\sigma}(t)$ and the transfer matrix $\hat{W}$ defined by $\mathbf{r}(t) = \hat{W}(t, t_1)\mathbf{r}(t_1)$. It satisfies the equation $d\hat{W}/dt = \hat{\sigma}\hat{W}$. The Liouville theorem $\mathrm{tr}\,\hat{\sigma} = 0$ means that $\det \hat{W} = 1$. The modulus $r(t)$ of the separation vector may be expressed via the positive symmetric matrix $\hat{W}^T\hat{W}$. The main result (Furstenberg and Kesten 1960; Oseledec, 1968) states that in almost every realization $\hat{\sigma}(t)$, the matrix $\frac{1}{t}\ln \hat{W}^T(t, 0)\hat{W}(t, 0)$ tends to a finite limit as $t \to \infty$. In particular, its eigenvectors tend to d fixed orthonormal eigenvectors $\mathbf{f}_i$. Geometrically, that

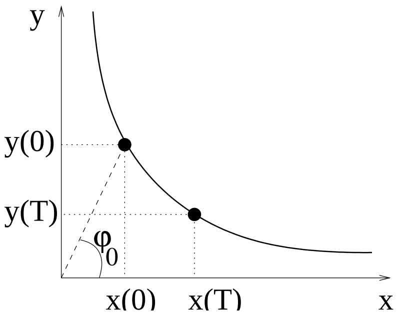


Figure 2: The distance of the point from the origin increases if the angle is less than $\varphi_0 = \arccos[1+\exp(2\lambda T)]^{-1/2} > \pi/4$. Note that for $\varphi = \varphi_0$ the initial and final points are symmetric relative to the diagonal: $x(0) = y(T)$ and $y(0) = x(T)$.

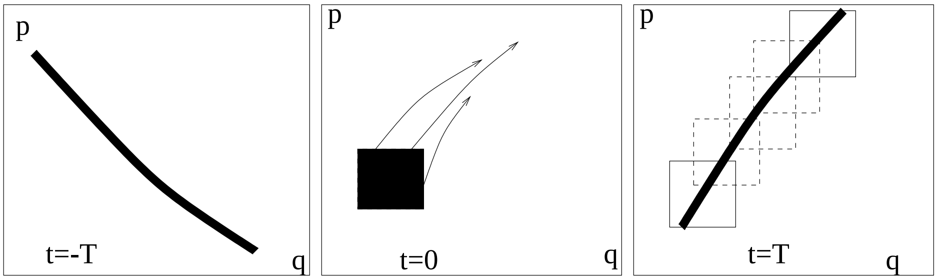
precisely means than an initial sphere evolves into an elongated ellipsoid at later times. The limiting eigenvalues

$$\lambda_i = \lim_{t\to\infty} t^{-1} \ln |\hat{W}\mathbf{f}_i| \tag{40}$$

define the so-called Lyapunov exponents. The sum of the exponents is zero due to the Liouville theorem so there exists at least one positive exponent which gives stretching. Therefore, as time increases, the ellipsoid is more and more elongated and it is less and less likely that the hierarchy of the ellipsoid axes will change. Mathematical lesson to learn is that multiplying N random matrices with unit determinant (recall that determinant is the product of eigenvalues), one generally gets some eigenvalues growing (and some decreasing) exponentially with N. It is also worth remembering that in a random flow there is always a probability for two trajectories to come closer. That probability decreases with time but it is finite for any finite time. In other words, majority of trajectories separate but some approach. The separating ones provide for the exponential growth of positive moments of the distance: $E(a) = \lim_{t\to\infty} t^{-1}\langle r^a(t)/r^a(0)\rangle > 0$ for $a > 0$. However, approaching trajectories have $r(t)$ decreasing, which guarantees that the moments with sufficiently negative a also grow. Mention without proof that $E(a)$ is a concave function, which evidently passes through zero, $E(0) = 0$. It must then have another zero which for isotropic random flow in d-dimensional space can be shown to be $a = -d$, see home exercise.

The probability to find a ball turning into an exponentially stretching ellipse thus goes to unity as time increases. The physical reason for it is that substantial deformation appears sooner or later. To reverse it, one needs to contract the long axis of the ellipse, that is the direction of contraction must be inside the narrow angle defined by the ellipse eccentricity, which is less likely than being outside the angle. Randomly oriented deformations on average continue to increase the eccentricity.

Armed with the understanding of the exponential stretching, we now return to the dynamical foundation of the second law of thermodynamics. We assume that our finite resolution does not allow us to distinguish between the states within some square in the phase space. That square is our "grain" in coarse-graining. In the figure below, one can see how such black square of initial conditions (at the central box) is stretched in one (unstable) direction and contracted in another (stable) direction so that it turns into a long narrow strip (left and right boxes). Later in time, our resolution is still restricted - rectangles in the right box show finite resolution (this is coarse-graining). Viewed with such resolution, our set of points occupies larger phase volume (i.e. corresponds to larger entropy) at $t = \pm T$ than at $t = 0$. Time reversibility of any particular trajectory in the phase space does not contradict the time-irreversible filling of the phase space by the set of trajectories considered with a finite resolution. By reversing time we exchange stable and unstable directions (i.e. those of contraction and expansion), but the fact of space filling persists. We see from the figure that the volume and entropy increase both forward and backward in time. To avoid misunderstanding, note that usual arguments that entropy growth provides for time arrow are such: if we already observed an evolution that produces a narrow strip then its time reversal is contraction into a ball; but if we consider a narrow strip as an initial condition, it is unlikely to observe a contraction because of the narrow angle mentioned above. Therefore, being shown two movies, one with stretching, another with contraction we conclude that with probability close (but not exactly equal!) to unity the first movie shows the true sequence of events, from the past to the future.



After the strip length reaches the scale of the velocity change (when one already cannot approximate the phase-space flow by a linear profile $\hat{\sigma} r$), strip starts to fold because rotation (which we can neglect for a ball but not for a long strip) is different at different parts of the strip. Still, however long, the strip continues locally the exponential stretching. Eventually, one can find the points from the initial ball everywhere which means that the flow is mixing, also called ergodic. Formal definition is that the flow is called ergodic in the domain if the trajectory of almost every point (except possibly a set of zero volume) passes arbitrarily close to every other point. An equivalent definition is that there are no finite-volume subsets of the domain invariant with respect to the flow except the domain itself. Ergodic flow on an energy surface in the phase space provides for a micro-canonical distribution (i.e. constant), since time averages are equivalent to the average over the surface. While we can prove ergodicity only for relatively simple systems, like the gas of hard spheres, we believe that it holds for most systems of sufficiently general nature (that vague notion can be make more precise by saying that the qualitative systems behavior is insensitive to small variations of its microscopic parameters).

When the density spreads, entropy grows (as the logarithm of the volume occupied). If initially our system was within the phase-space volume ϵ^{6N}, then its density was $\rho_0 = \epsilon^{-6N}$ inside and zero outside. After stretching to some larger volume $A\epsilon^{6N}$ the entropy $S = -\int \rho \ln \rho d\mathbf{x}$ has increased by $\ln A$. The Lyapunov exponent determines the rate of the entropy growth.

Two concluding remarks are in order. First, the notion of an exponential separation of trajectories put an end to the old dream of Laplace to be able to predict the future if only all coordinates and momenta are given. Even if we were able to measure all relevant phase-space initial data, we can do it only with a finite precision ϵ. However small is the indeterminacy in the data, it is amplified exponentially with time so that eventually $\epsilon \exp(\lambda T)$ is large and we cannot predict the outcome. Mathematically speaking, limits $\epsilon \to 0$ and $T \to \infty$ do not commute. Second, the above arguments did not use the usual mantra of thermodynamic limit, which means that even the systems with a small number of degrees of freedom need statistics for their description at long times if their dynamics has a positive Lyapunov exponent (which is generic) - this is sometimes called *dynamical chaos*.

3.2 Information theory approach

Information is physical (Landauer)

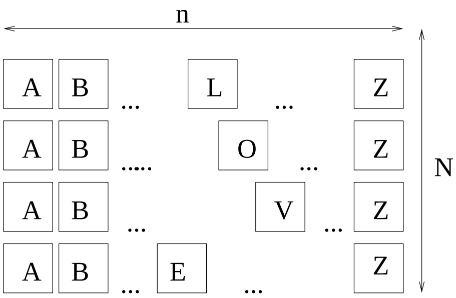
Here I briefly re-tell the story of statistical physics using a different language. It will let us see entropy in a new light. An advantage of using different formulations is that it helps to understand things better and triggers different intuition in different people.

Consider first a simple problem in which we are faced with a choice among n equal possibilities (say, in which of n boxes a candy is hidden). How much we need to know to get the candy? Let us denote the missing information by $I(n)$. Clearly, the information is an increasing function of n and $I(1) = 0$. If we have several independent problems then information must be additive. For example, consider each box to have m compartments: $I(nm) = I(n) + I(m)$. Now, we can write (Shannon, 1948)

$$I(n) = I(e) \ln n = k \ln n \tag{41}$$

That it must be a logarithm is clear also from obtaining the missing information by asking the sequence of questions in which half we find the box with the candy, one then needs $\log_2 n$ of such questions and respective one-bit answers. We can easily generalize the definition (41) for non-integer rational numbers by $I(n/l) = I(n) - I(l)$ and for all positive real numbers by considering limits of the series and using monotonicity. So the single number of the lucky box with the candy brings the information $k \ln n$.

We used to think of information received through words and symbols. If we have an alphabet with n symbols then every symbol we receive is a choice out of n and brings the information $k \ln n$. If symbols come independently then the message of the length N can potentially be one of n^N possibilities so that it brings the information $kN \ln n$. If all the 25 letters of the English alphabet were used with the same frequency then the word "love" would bring the information equal to $4k \ln 25$ or $4 \log_2 25$ bits. Here and below we assume that the receiver has no other prior knowledge on subjects like correlations between letters (for instance, everyone who knows English, can infer that there is only one four-letter word which starts with “lov...” so the last letter brings zero information for such people).



In reality though every letter brings even less information than $k\ln 25$ since we know that letters are used with different frequencies. Indeed, consider the situation when there is a probability w_i assigned to each letter (or box) $i = 1, \ldots, n$. It is then clear that different letters bring different information. When there is randomness, we evaluate the *average* information per symbol by repeating our choice, say, N times. As $N \to \infty$ we know that candy in the i-th box in Nw_i cases but we do not know the order in which different possibilities appear. Total number of orders is $N!/\,\Pi_i(Nw_i)!$ and the information that we obtained from N symbols is

$$I_N = k\ln\Big(N!/\,\Pi_i(Nw_i)!\Big) \approx -Nk\sum_i w_i \ln w_i + O(lnN) \ . \qquad (42)$$

The missing information per symbol in the language coincides with the entropy (26):

$$I(w_1 \ldots w_n) = \lim_{N\to\infty} I_N/N = -k\sum_{i=1}^{n} w_i \ln w_i \ . \qquad (43)$$

Incidentally for English language the information per symbol is

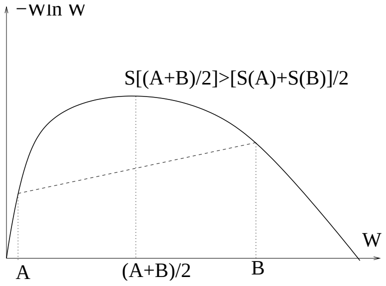
$$-\sum_{i=a}^{z} w_i \log_2 w_i \approx 4.11\,\mathrm{bits} \ .$$

The information (43) is zero for delta-distribution $w_i = \delta_{ij}$; it is generally less than the information (41) and coincides with it only for equal probabilities, $w_i = 1/n$, when the entropy is maximum. Indeed, equal probabilities we ascribe when there is no extra information, i.e. in a state of maximum ignorance. In this state, we get maximum information per symbol; any prior knowledge can reduce the information. Mathematically, the property

$$I(1/n, \ldots, 1/n) \geq I(w_1 \ldots w_n) \qquad (44)$$

is called convexity. It follows from the fact that the function of a single variable $s(w) = -w\ln w$ is strictly downward convex (concave) since its second derivative, $-1/w$, is everywhere negative for positive w. For any concave function, the average over the set of points w_i is less or equal to the function at the average value (so-called Jensen inequality):

$$\frac{1}{n}\sum_{i=1}^{n} s\left(w_i\right) \le s\left(\frac{1}{n}\sum_{i=1}^{n} w_i\right) \ . \tag{45}$$



From here one gets the entropy inequality:

$$I(w_1 \dots w_n) = \sum_{i=1}^{n} s\left(w_i\right) \le ns\left(\frac{1}{n}\sum_{i=1}^{n} w_i\right) = ns\left(\frac{1}{n}\right) = I\left(\frac{1}{n}, \dots, \frac{1}{n}\right) \ . \tag{46}$$

The relation (45) can be proven for any concave function. Indeed, the concavity condition states that the linear interpolation between two points a, b lies everywhere below the function graph: $s(\lambda a + b - \lambda b) \ge \lambda s(a) + (1-\lambda)s(b)$ for any $\lambda \in [0,1]$, see the Figure. For $\lambda = 1/2$ it corresponds to (45) for $n = 2$. To get from $n = 2$ to arbitrary n we use induction. For that end, we choose $\lambda = (n-1)/n$, $a = (n-1)^{-1}\sum_{i=1}^{n-1} w_i$ and $b = w_n$ to see that

$$\begin{aligned} s\left(\frac{1}{n}\sum_{i=1}^{n} w_i\right) &= s\left(\frac{n-1}{n}(n-1)^{-1}\sum_{i=1}^{n-1} w_i + \frac{w_n}{n}\right) \\ &\ge \frac{n-1}{n} s\left((n-1)^{-1}\sum_{i=1}^{n-1} w_i\right) + \frac{1}{n} s\left(w_n\right) \\ &\ge \frac{1}{n}\sum_{i=1}^{n-1} s\left(w_i\right) + \frac{1}{n} s\left(w_n\right) = \frac{1}{n}\sum_{i=1}^{n} s\left(w_i\right) \ . \end{aligned} \tag{47}$$

In the last line we used the truth of (45) for $n-1$ to prove it for n.

Note that when $n \to \infty$ then (41) diverges while (43) may well be finite. We can generalize (43) for a continuous distribution by dividing into cells (that is considering a limit of discrete points). Here, different choices of variables to define equal cells give different definitions of information. It is in such a choice that physics (or other specific knowledge) enters. Physics (quantum mechanics) requires that for Hamiltonian system the equal volumes in phase space contain equal number of states, so the measure is uniform in canonical coordinates; we then write the missing information in terms of the phase space density, which may also depend on time:

$$I(t) = -\int \rho(p,q,t)\ln[\rho(p,q,t)]\,dpdq\ . \tag{48}$$

It is maximal for the uniform distribution $\rho = 1/\Gamma$, $I = \ln\Gamma$.

If the density of the discrete points in the continuous limit is inhomogeneous, say $m(\mathbf{x})$, then the proper generalization is

$$I(t) = -\int \rho(\mathbf{x})\ln[\rho(\mathbf{x})/m(\mathbf{x})]\,d\mathbf{x}\ .$$

It is invariant with respect to an arbitrary change of variables $\mathbf{x} \to \mathbf{y}(\mathbf{x})$ since $\rho(\mathbf{y})d\mathbf{y} = \rho(\mathbf{x})d\mathbf{x}$ and $m(\mathbf{y})d\mathbf{y} = m(\mathbf{x})d\mathbf{x}$ while (48) was invariant only with respect to canonical transformations (including a time evolution according to a Hamiltonian dynamics) that conserve the element of the phase-space volume. If we introduce the normalized distribution of points $\rho'(\mathbf{x}) = m(\mathbf{x})/\Gamma$, then

$$I(t) = \ln V - \int \rho(\mathbf{x})\ln[\rho(\mathbf{x})/\rho'(\mathbf{x})]\,d\mathbf{x}\ . \tag{49}$$

The last term in (49) turns into zero when ρ and ρ' coincide and thus presents some measure of the difference between the distributions.

Mention briefly the application of entropy in communication theory. Inequality (44) means that a communication channel transmitting bits (ones and zeros) on average can transmit no more than one unit of the information (43) per symbol. In other words, $\sum_{i=a}^{z} w_i \log_2 w_i$ gives the minimum number of bits per symbol needed to transmit the ensemble of messages. We can say that the information content of a symbol number i is $\log_2(1/w_i)$, while the entropy is the mean information content per symbol. Note that less probable symbols have larger information content, but they happen more rarely. The mean information content for a given letter, $-w\ln w$, is maximal for $w = 1/e$. Different probability of letters suggests a way of signal compression by coding common letters by short sequences and infrequent letters by more lengthy combinations - lossless compressions like zip, gz and gif work this way (you may find it interesting to know that jpeg, mpeg, mp3 and telephone use lossy compression which removes information presumed to be unimportant for humans).

Apart from restrictions imposed by the statistics of symbols to be transferred, one also wish to characterize the quality of the channel. Note that in this context one can view measurements as messages about the value of the quantity we measure. Here, the message (measurement) A we receive gives the information about the event (quantity) B. That information must be related to the fact that having observed A increases the probability to have B comparing to the unconditional probability to observe B:

$$I(A,B)=\ln[P(B|A)/P(B)] \ .$$

Here $P(B|A)$ is the so-called conditional probability (of B in the presence of A). The conditional probability is related to the joint probability $P(A,B)$ by the evident formula $P(A,B)=P(B|A)P(A)$, which allows one to write the information in a symmetric form

$$I(A,B)=\ln\left[\frac{[P(B,A)}{P(A)P(B)}\right] \ . \tag{50}$$

When A and B are independent then the conditional probability is independent of A and information is zero. When they are dependent, $P(B,A)\geq P(A)P(B)$, so that that the information is always positive.

If one is just interested in the channel as specified by $P(B|A)$ then one maximizes $I(A,B)$ over all choices of the source statistics $P(B)$ and call it the channel capacity. Alternatively, one may wish to know how much information about B one obtains *on average* by measuring A. Summing over all possible $B_1,\ldots,B_n$ and $A_1,\ldots,A_m$ we obtain Shannon's "mutual information" used to evaluate the quality of communication systems (or measurements)

$$\begin{aligned} I(A,B)&=\sum_{i=1}^{m}\sum_{j=1}^{n}P(A_i,B_j)\ln[P(B_j|A_i)/P(B_j)] \\ \to I(Z,Y)&=\int dzdyp(z,y)\ln\left[\frac{p(z|y)}{p(y)}\right]=\int dzdy\,p(z,y)\ln\left[\frac{p(z,y)}{p(z)p(y)}\right] \ . \end{aligned} \tag{51}$$

Here we used $p(z,y)=p(z|y)p(y)$ - the probability to get y,z is the probability to get y times the probability to get z for this y. Note that (51) is the particular case of multidimensional (49), where one takes $\mathbf{x}=(y,z)$, $\rho'=p(z)p(y)$, that is mutual information measures the difference between the true joint distribution and the distribution taken as if the quantities were statistically independent. It is straightforward to generalize it from the pair to many quantities.

You probably noticed that (41,50) corresponds to the microcanonical description (16) giving information/entropy as a logarithm of the number of states, while (43,49,51) corresponds to the canonical description (26) giving it as an average. An advantage of Shannon entropy (43,49,51) is that it is defined for arbitrary distribution, not necessarily equilibrium. One can go further and define a free energy for any system in a contact with a thermostat having temperature T as $F(\rho) = E(\rho) - TS(\rho)$, even when the distribution of the system itself is not equilibrium.

Mutual information also allows to understand the second law of thermodynamics from a different perspective. Boltzmann considered the thermodynamic entropy of the gas as a the sum of entropies of different particles $\sum S(p_i, q_i)$, neglecting their correlations, i.e. the mutual information $\sum_i S(p_i, q_i) - S(p_1 \ldots p_n, q_1, \ldots q_n) = I(p_1 \ldots p_n, q_1, \ldots q_n)$. That allowed him to establish H-theorem, that is the growth of the thermodynamic (uncorrelated) entropy. Since the Liouville theorem guarantees that the true entropy $S(p_1 \ldots p_n, q_1, \ldots q_n)$ does not change upon evolution, then the increase of the uncorrelated part must be compensated by the increase of the mutual information. In other words, one can replace the usual second law of thermodynamics by the law of conservation of the total entropy (or information), where the increase in the thermodynamic (uncorrelated) entropy is exactly compensated by the increase in correlations between particles expressed by the mutual information. The usual second law then results simply from our renunciation of all correlation knowledge, and not from any intrinsic behavior of dynamical systems.

So far, we defined information via the distribution. Now, we want to use the idea of information to get the distribution. Statistical mechanics is a systematic way of guessing, making use of incomplete information. The main problem is how to get the best guess for the probability distribution $\rho(p, q, t)$ based on any given information presented as $\langle R_j(p, q, t) \rangle = r_j$, i.e. as the expectation (mean) values of some dynamical quantities. Our distribution must contain *the whole truth* (i.e. all the given information) and *nothing but the truth* that is it must maximize the missing information I. This is to provide for the widest set of possibilities for future use, compatible with the existing information. Looking for the maximum of

$$I-\sum_j \lambda_j\langle R_j(p,q,t)\rangle=\int \rho(p,q,t)\{\ln[\rho(p,q,t)]-\sum_j \lambda_j R_j(p,q,t)\}\,dpdq\ ,$$

we obtain the distribution

$$\rho(p,q,t)=Z^{-1}\exp\Big[-\sum_j \lambda_j R_j(p,q,t)\Big]\ , \tag{52}$$

where the normalization factor

$$Z(\lambda_i)=\int \exp\Big[-\sum_j \lambda_j R_j(p,q,t)\Big]\,dpdq\ ,$$

can be expressed via the measured quantities by using

$$\frac{\partial \ln Z}{\partial \lambda_i}=-r_i\ . \tag{53}$$

For example, consider our initial "candy-in-the-box" problem (think of an impurity atom in a lattice if you prefer physics to candies). Let us denote the number of the box with the candy j. Different attempts give different j (for impurity, think of X-ray with wavenumber k scattering on the lattice) but on average after many attempts we find, say, $\langle\cos(kj)\rangle=0.3$. Then

$$\rho(j)=Z^{-1}(\lambda)\exp[-\lambda\cos(kj)]$$

$$Z(\lambda)=\sum_{j=1}^{n}\exp[\lambda\cos(kj)]\,,\quad \langle\cos(kj)\rangle=d\log Z/d\lambda=0.3\ .$$

We can explicitly solve this for $k\ll 1\ll kn$ when one can approximate the sum by the integral so that $Z(\lambda)\approx nI_0(\lambda)$ where I_0 is the modified Bessel function. Equation $I_0'(\lambda)=0.3I_0(\lambda)$ has an approximate solution $\lambda\approx 0.63$.

Note in passing that the set of equations (53) may be self-contradictory or insufficient so that the data do not allow to define the distribution or allow it non-uniquely. If, however, the solution exists then (48,52) define the missing information $I\{r_i\}$ which is analogous to thermodynamic entropy as a function of (measurable) macroscopic parameters. It is clear that I have a tendency to increase whenever a constraint is removed (when we measure less quantities R_i).

If we know the given information at some time t_1 and want to make guesses about some other time t_2 then our information generally gets less relevant as the distance $|t_1 - t_2|$ increases. In the particular case of guessing the distribution in the phase space, the mechanism of loosing information is due to separation of trajectories described in Sect. 3. Indeed, if we know that at t_1 the system was in some region of the phase space, the set of trajectories started at t_1 from this region generally fills larger and larger regions as $|t_1 - t_2|$ increases. Therefore, missing information (i.e. entropy) increases with $|t_1 - t_2|$. Note that it works both into the future and into the past. Information approach allows one to see clearly that there is really no contradiction between the reversibility of equations of motion and the growth of entropy.

Yet there is one class of quantities where information does not age. They are integrals of motion. A situation in which only integrals of motion are known is called equilibrium. The distribution (52) takes the canonical form (36,37) in equilibrium. On the other hand, taking micro-canonical as constant over the constant-energy surface corresponds to the same approach of not adding any additional information to what is known (energy).

From the information point of view, the statement that systems approach equilibrium is equivalent to saying that all information is forgotten except the integrals of motion. If, however, we possess the information about averages of quantities that are not integrals of motion and those averages do not coincide with their equilibrium values then the distribution (52) deviates from equilibrium. Examples are currents, velocity or temperature gradients like considered in kinetics.

The concept of entropy as missing information[8] allows one to understand that entropy does not really decrease in the system with Maxwell demon or any other information-processing device (indeed, if at the beginning one has an information on position or velocity of any molecule, then the entropy was less by this amount from the start; after using and processing the information the entropy can only increase). Consider, for instance, a particle in the box. If we know that it is in one half then entropy (the logarithm of *available* states) is $\ln(V/2)$. That also teaches us that information has thermodynamic (energetic) value: by placing a piston at the half of the box and allowing particle to hit and move it we can get the work $T\Delta S = T \ln 2$ done. On the other hand, the law of energy conservation tells that to get such an

---

[8] that entropy is not a property of the system but of our knowledge about the system

information one must make a measurement whose minimum energetic cost is $T\Delta S = T \ln 2$ (that was considered by Szilard in 1929 who also introduced "bit" as a unit of information). Making measurement R one changes the distribution from $\rho(x)$ to (generally non-equilibrium) $\rho(x|R)$, which decreases the entropy of the system by the mutual information (51):

$$\begin{aligned} S(x|R) - S(x) = -\int \rho(x|R) \ln \rho(x|R)\, dxdR \\ + \int \rho(x) \ln \rho(x)\, dx = \int \rho(x,R) \ln[\rho(x,R)/\rho(x)\rho(R)]\, dxdR\ . \end{aligned}$$

If such measurement does not change energy (like the knowledge in which half of the box the particles is), the entropy decrease increases the (non-equilibrium) free energy, so that the minimal work to perform such a measurement is $F(\rho(x|R)) - F(\rho(x))$.

More details can be found in Katz, Sects. 2-5, Sethna Sect. 5.3 and Kardar I, Problem 2.6.