PART II. FOUNDATIONS OF STAT. MECHANICS

§6. Principle of Maximum Entropy.

We now want to develop a completely general approach to stat. physics of macroscopic systems. (systems with many degrees of freedom, many particles)

Like any quantum mechanical system (& all natural systems are quantum mechanical) such a system is basically something that can have many states [collective] (quantum states – call them microstates), each with some probability.

[Measuring them is practically impossible, but we assume, by force of pure thought, that in a gedanken experiment, we can measure such microstates precisely & obtain each distinct result with a certain probability]

Let's enumerate these states (they are discrete in QM!)

$\alpha = 1, 2, 3, \dots, \Omega \gg\gg\gg 1$ (total # of states)

$p_1, p_2, p_3, \dots, p_\Omega$ their probabilities, $\sum_{\alpha=1}^{\Omega} p_\alpha = 1$

$E_1, E_2, E_3, \dots, E_\Omega$ their energies [to fix microstates exactly]

(and we can specify other quantities associated with the microstates because of course many diff. microstates can have the same energy, so E_α are not all different)

e.g. for particles in ideal gas, we had velocities (or momenta)
α = {a state with some fixed # of particles with each momentum}

In principle, if we could solve QM for our system, it would tell us what $\{E_\alpha\}$ are (or any other quantities). If we also knew $\{p_\alpha\}$, then we could calculate average quantities, e.g. mean energy:

$$U = \sum_\alpha p_\alpha E_\alpha = \langle E_\alpha \rangle$$

- and other mean quantities that characterise the macroscopic state of the system.

While I have not yet defined them precisely (e.g. what is pressure in terms of these p_α's and E_α's?), I will do eventually - for now I just want it to be clear that it would be useful to devise a way to calculate probabilities p_α for each quantum state the system can have and the corresponding energies $\{E_\alpha\}$ etc.

- It is obvious that I can't make any progress on calculating $\{E_\alpha\}$ w/o specifying precisely what my system is made of and how - so this is a job for microphysical [in general, quantum] theory. Usually some model is needed for that (e.g. classical ideal gas or some model of a real/quantum gas or crystal or whatever).
- The amazing thing is that we can determine the probabilities p_α completely generally as functions of E_α!

6.1 Fundamental Postulate

To do this, let's start from the position of complete

Ignorance about the system.

If we know nothing of it, we might assume that all the microstates are equally likely (in equilibrium)

$$\boxed{p_\alpha = \frac{1}{\Omega}}$$

This of course does not really follow from anywhere, but it seems sensible as we might even give it the status of a postulate:

Fundamental postulate of stat. mech. [Boltzmann]

(aka Equal a priori probability postulate)

For an isolated system in equilibrium, all microstates are equally likely

↕ means p_α are not changing, it's all sorted itself out into a steady state (statistical!)

means it is not in contact with anything (which, incidentally, is consistent with us not knowing anything about it)

Now imagine we do know something — which in practice means we have measured something, e.g. the average energy

(*) $\sum_\alpha p_\alpha E_\alpha = U$ — this is not necessarily consistent with $p_\alpha = 1/\Omega$ for a given set of $\{E_\alpha\}$. But it is not a terribly stringent constraint because after all there are very many p_α's and only one equation (*) that they have to satisfy.

So we'd like to adjust the distribution a bit to accommodate the constraint - but not too drastically, leave it as close to equal probabilities as allowed by the little knowledge we possess. [e.g. would be silly to think there is only one microstate with $p_2 = 1$ and $E_2 = U$]

If we had further knowledge, we'd make further adjustments — we clearly need a systematic way of doing this.

6.2 Entropy

Let's formalise this process so: (to measure uncertainty) we will introduce some function (called entropy)

$$S = S(p_1, p_2, \cdots, p_\Omega)$$

which characterises any given set of probabilities $\{p_\Omega\}$ as which must be chosen in such a way that maximising S subject to a set of constraints (representing what we know) gives us the probability distribution corresponding to maximum uncertainty given those constraints.

[NB: we don't want to put in more info than we really possess!]

(maximise it power!)

In order to be suitable, S must satisfy certain requirements:

1) $S(p_1, p_2, \cdots, p_\Omega) < S\left(\frac{1}{\Omega}, \cdots, \frac{1}{\Omega}\right) \equiv S_\Omega$ ← special case of S

(max. uncertainty ⇐⇒ equal probabilities - fund. postulate)

2) If $\Omega' > \Omega$, then $S_{\Omega'} > S_\Omega$

(more possibilities ⇒ more uncertainty)

3) $S(p_1, \cdots, p_\Omega)$ is symmetric wrt permutation of $\{p_\Omega\}$

(does not matter in what order we list the microstates)

4) S is independent of how we count the states. (additive law)

Namely: let us divide microstates into groups:

$\alpha = 1, \dots, m_1, m_1+1, \dots, m_1+m_2, \dots, \dots \sum_{i=1}^{M} m_i = \Omega$ students

$i = 1 \quad 2 \quad \dots \quad M$ colleges

$w_1 \quad w_2 \quad \dots \quad w_M$ probabilities of these groups

So we now have the probability distribution of groups $\{w_1 \dots w_M\}$ and a probability distribution of microstates within each group:

for group i, $\{p_1^{(i)}, \dots, p_{m_i}^{(i)}\}$. NB: $\sum_{i=1}^{M} w_i = 1$, $\sum_{k=1}^{m_i} p_k^{(i)} = 1$

Consider some microstate α from the original set. W/o loss of generality, suppose it's in the first set ($i=1$) and it's in fact #1 (the order in which we list them does not matter!). Then

$$p_1 = w_1 \cdot p_1^{(1)} \quad \Longrightarrow \quad p_1^{(1)} = \frac{p_1}{w_1}$$

(w_1: prob. to pick group 1; $p_1^{(1)}$: prob. to pick state 1 in that group)

Similarly, for others:

$$p_k^{(i)} = \frac{p_\alpha}{w_i}$$, where i is the group #, α is the state in the original set that #k in group i corresponds.

We want entropy to be <u>additive</u> in the following sense:

$$S(p_1 \dots p_\Omega) = S(w_1 \dots w_M) + w_1 S\left(p_1^{(1)} \dots p_{m_1}^{(1)}\right) + w_2 S\left(p_1^{(2)} \dots p_{m_2}^{(2)}\right) + \dots$$

$S(w_1 \dots w_M)$: uncertainty in the distr. of groups.

w_2: prob. of a particular group

$S\left(p_1^{(2)} \dots p_{m_2}^{(2)}\right)$: uncertainty in the distr. within the group

$$= S(w_1 \dots w_M) + w_1 S\left(\frac{p_1}{w_1}, \dots, \frac{p_{m_1}}{w_1}\right) + w_2 S\left(\frac{p_{m_1+1}}{w_2}, \dots, \frac{p_{m_1+m_2}}{w_2}\right) + \dots \quad (1)$$

Consider now a <u>special case</u>: suppose $p_\alpha = \frac{1}{\Omega}$.

Then $w_i = \frac{m_i}{\Omega}$ and $p_k^{(i)} = \frac{p_\alpha}{w_i} = \frac{1}{w_i \Omega} = \frac{1}{m_i}$

$$S(p_1 \dots p_\Omega) = S_\Omega = S(w_1 \dots w_M) + \sum_{i=1}^{M} w_i \underbrace{S\left(\frac{1}{m_i}, \dots, \frac{1}{m_i}\right)}_{\| \; S_{m_i}}$$

So, we have

$$S(w_1, \dots, w_M) = S_\Omega - \sum_{i=1}^{M} w_i S_{m_i} \quad (2)$$

Now consider <u>an even more special case</u>:

Suppose all $m_i = m$ are the same. Then $w_i = \frac{m_i}{\Omega} = \frac{m}{\Omega} = \frac{1}{N}$

From (2),

$$S_M = \underbrace{S_{mM}}_{\Omega} - S_m \quad (3)$$

(3) is a functional equation : in other words,
if $S_n = f(n)$, then $f(nm) = f(n) + f(m)$.
It is possible to prove that the only monotonic (increasing) solution of this is

$$f(n) = \text{const} \cdot \ln n$$

or $$\boxed{S_\Omega = k_B \ln \Omega}$$ Boltzmann formula (on his grave!)

because we require $S_{n'} > S_n$ if $n' > n$ see p-2

for historical reasons k_B chosen to be the B. constant – we'll see why soon, but of course it does not matter.

This will be needed to have the usual units of temperature

Proof:

If $f(mn) = f(m) + f(n)$,
then
$$f(n^k) = k f(n)$$

Let u, v be integers > 1. Then we can always find m and n (can arb. large) such that

$$\frac{m}{n} < \frac{\ln v}{\ln u} < \frac{m+1}{n} \Leftrightarrow u^m < v^n < u^{m+1}$$

Since f is monotonically increasing,

$$f(u^m) < f(v^n) < f(u^{m+1})$$

$$m f(u) < n f(v) < (m+1) f(u)$$

$$\frac{m}{n} < \frac{f(v)}{f(u)} < \frac{m+1}{n}$$

$$\left| \frac{f(v)}{f(u)} - \frac{\ln v}{\ln u} \right| < \frac{1}{n} \Rightarrow \left| \frac{f(v)}{\ln(v)} - \frac{f(u)}{\ln(u)} \right| < \frac{1}{n} \frac{f(u)}{\ln v}$$

can be made arb. small (n arb. large!)

So $\frac{f(v)}{\ln v} = \frac{f(u)}{\ln u} = \text{const}$ q.e.d.

From Boltzmann formula and (2), we get

$$S(w_1,\dots,w_M) = k_B \ln \Omega - \sum_{i=1}^{M} w_i\, k_B \ln m_i =$$

$$= -k_B \sum_{i=1}^{M} w_i \ln \frac{m_i}{\Omega} = -k_B \sum_{i=1}^{M} w_i \ln w_i$$

But remember that $\{m_i\}$ were chosen in a general way, subject only to $\sum_{i=1}^{M} m_i = \Omega$, or $\sum_{i=1}^{M} w_i = 1$.

Thus, with equal validity,

$$\boxed{S(p_1,\dots,p_\Omega) = -k_B \sum_\alpha p_\alpha \ln p_\alpha} = k_B \langle -\ln p_\alpha \rangle$$

Gibbs formula.

Note. Technically, we only obtained the above for probabilities that are rational numbers ($w_i = \frac{m_i}{\Omega}$). But that's OK because even if p_α are irrational, they can be approximated arbitrarily well by rational numbers as S is continuous (can assume that)

So, our strategy for finding $\{p_\alpha\}$ will be to maximize S_x (and we can elevate this to the status of a postulate) subject to constraints.

E.g. 1) $\sum_\alpha p_\alpha = 1$ always must be in place (condition (1) on S (p. 52))

→ If we impose nothing else, must get $p_\alpha = \frac{1}{\Omega}$ (*)

so equal a priori prob. postulate will follow

2) $\sum_\alpha p_\alpha E_\alpha = U$ mean energy $\Longrightarrow$?

L.7 ended here.

These Comments on p. 61

6.3. Shannon entropy

non-examinable

Note: Gibbs first came up with the idea of determining p_α's by maximizing S, but did not explain why. Claude Shannon (1948) worked it out in the context of info. theory.

Entropy is directly connected to Shannon's function, which measures "information content" of a probability distribution.

Suppose we are transmitting messages consisting of characters

$\alpha = 1, \dots, \Omega$

$p_1 \dots p_\Omega$ – probabilities with which these characters may occur.

Suppose our message has N characters.

If all characters occur with the same probability ($p_\alpha = \frac{1}{\Omega}$) then the number of all possible messages is

$$n = \Omega^N$$

and the number of bits of info needed to transmit a message (i.e. distinguish between these possibilities) is

$$\log_2 n = N \log_2 \Omega$$

[with this many bits there are $2^{\log_2 n} = n$ distinct binary sequences]

If p_α's are not the same, some symbols are more frequent than others. If $N \gg \Omega$, there will be approximately

$N_\alpha = N p_\alpha$ occurrences of symbol α in a message.

So, the # of typical messages is

$$n = \frac{N!}{N_1! N_2! \dots N_\Omega!}$$

$N!$ ← # of rearrangements of all symbols

↑ #'s of rearrangements of identical symbols (does not change the message)

Then the number of binary bits required is

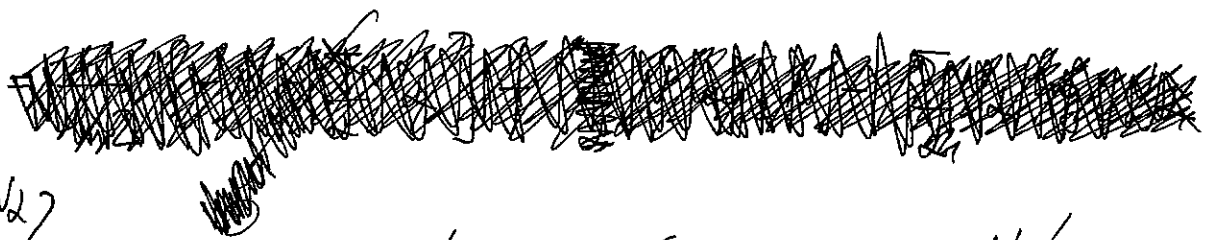
$$\log_2 n = \log_2 N! - \sum_\alpha \log_2 N_\alpha !$$

Stirling's formula:

$$\ln N! \approx N \ln N - N \text{ as } N \to \infty$$

$\log_2 = \frac{\ln}{\ln 2}$, so we get

$$\log_2 n \approx \frac{1}{\ln 2}\left[N \ln N - N\right] - \sum_\alpha \frac{1}{\ln 2}\left[N_\alpha \ln N_\alpha - N_\alpha\right]$$



use $N = \sum_\alpha N_\alpha$

$$= \sum_\alpha \left(N_\alpha \log_2 N - \frac{N_\alpha}{\ln 2}\right) - \sum_\alpha \left(N_\alpha \log_2 N_\alpha - \frac{N_\alpha}{\ln 2}\right) =$$

$$= -\sum_\alpha N_\alpha \log_2 \frac{N_\alpha}{N} = -N \sum_\alpha p_\alpha \log_2 p_\alpha$$

→ see Bell Systems Technical Journal 27, 379 (1948)

Shannon's theorem states rigorously that this is indeed the minimum # of bit necessary to ensure that the fraction of errors in a message vanishes as $N \to \infty$

The difference per character in a message between $\log_2 n$ and its max. value $+ N \log_2 \Omega$ is the info content of the distribution:

$$I(p_1 \cdots p_\Omega) = \frac{N \log_2 \Omega + N \sum_\alpha p_\alpha \log_2 p_\alpha}{N}$$

$$= \log_2 \Omega + \sum_\alpha p_\alpha \log_2 p_\alpha$$

$I = 0$ if $p_\alpha = 1/\Omega$. [Constants ($\frac{1}{\ln 2}$ rather than k_B do not, of course matter)]

Minimising this is the same as maximising entropy.

6.4 Lagrange Multipliers [Mathematical digression]

How do we maximise a function $S(p_1, \dots, p_\Omega)$ subject to some constraint in the form

$$F(p_1, \dots, p_\Omega) = 0$$

↳ e.g. $F = \sum_\alpha p_\alpha - 1$ or $\sum_\alpha p_\alpha E_\alpha - U$.

At the point of maximum,

$$dS = \frac{\partial S}{\partial p_1} dp_1 + \dots + \frac{\partial S}{\partial p_\Omega} dp_\Omega = 0$$

but $\{dp_\alpha\}$ are not independent because we are only allowed to change $\{p_\alpha\}$ subject to them remaining on the surface $F(p_1, \dots, p_\Omega) = 0$. So F cannot change:

$$dF = \frac{\partial F}{\partial p_1} dp_1 + \dots + \frac{\partial F}{\partial p_\Omega} dp_\Omega = 0$$

From this can calculate one of the dp_α's in terms of others:

w.l.o.g., $$dp_\Omega = -\left[\frac{\partial F/\partial p_1}{\partial F/\partial p_\Omega} dp_1 + \dots + \frac{\partial F/\partial p_{\Omega-1}}{\partial F/\partial p_\Omega} dp_{\Omega-1} \right]$$

Then, substituting into dS:

$$dS = \left(\frac{\partial S}{\partial p_1} - \frac{\partial S}{\partial p_\Omega} \frac{1}{\partial F/\partial p_\Omega} \frac{\partial F}{\partial p_1} \right) dp_1 +$$

$$+ \dots \left(\frac{\partial S}{\partial p_{\Omega-1}} - \frac{\partial S}{\partial p_\Omega} \frac{1}{\partial F/\partial p_\Omega} \frac{\partial F}{\partial p_{\Omega-1}} \right) dp_{\Omega-1} = 0$$

$$\frac{\partial S}{\partial p_\Omega} \frac{1}{\partial F/\partial p_\Omega} = \lambda$$

All $dp_1 \dots dp_{\Omega-1}$ are now independent, so this implies

$$\frac{\partial S}{\partial p_\alpha} - \lambda \frac{\partial F}{\partial p_\alpha} = 0 \quad , \alpha = 1, \dots, \Omega - 1$$

and, by definition of λ,

$$\frac{\partial S}{\partial p_\Omega} - \lambda \frac{\partial F}{\partial p_\Omega} = 0 .$$

Thus, we now have $\Omega+1$ equations for $\Omega+1$ variables $p_1, \dots, p_\Omega, \lambda$: ← Lagrange multiplier

$$\begin{cases} \dfrac{\partial S}{\partial p_\alpha} - \lambda \dfrac{\partial F}{\partial p_\alpha} = 0 \quad , \alpha = 1, \dots, \Omega \\ \text{and } F(p_1, \dots, p_\Omega) = 0 \end{cases}$$

(wrt $p_1 \dots p_\Omega, \lambda$)

but we obtain the same exact equations if we simply maximise $S - \lambda F$ with no constraints ! Indeed, ← implies

$$d(S - \lambda F) = \sum_\alpha \left(\frac{\partial S}{\partial p_\alpha} - \lambda \frac{\partial F}{\partial p_\alpha} \right) dp_\alpha - F \, d\lambda = 0$$

If there are several constraints

$$F_i(p_1 \dots p_\Omega) = 0 \quad i = 1, \dots, m$$

the above procedure generalizes trivially to maximizing

$$S - \sum_i \lambda_i F_i \quad \text{wrt} \quad p_1 \dots p_\Omega, \lambda_1 \dots \lambda_m$$

involving as many Lagrange multipliers as there are constraints (obviously, we must have $m < \Omega$, but in practice you will never have enough info to constrain every state).

6.5 Equal a priori probabilities

Let's check that this procedure produces the right answer for an isolated system (= no info):

maximise $S(p_1 \dots p_\Omega)$ subject to $\sum_\alpha p_\alpha = 1$

$$S - \lambda\left(\sum_\alpha p_\alpha - 1\right) \to \max$$

$$\underbrace{-k_B \sum_\alpha dp_\alpha \left(\ln p_\alpha + p_\alpha \cdot \frac{1}{p_\alpha}\right) - \lambda \sum_\alpha dp_\alpha}_{\| \; -\sum_\alpha dp_\alpha (k_B \ln p_\alpha + k_B + \lambda)} - d\lambda \underbrace{\left(\sum_\alpha p_\alpha - 1\right)}_{\| \; 0} = 0$$

$$\Downarrow$$

$$\to p_\alpha = e^{-\left(1 + \frac{\lambda}{k_B}\right)} = \text{const} \to \Omega \cdot p_\alpha = 1$$

$$\Downarrow$$

$$\boxed{p_\alpha = \frac{1}{\Omega}}$$

q.e.d.

So we have proven that $S(p_1 \dots p_\Omega) < S_\Omega$

Ok, so our algorithm works.

Now we are in business and can work out absolutely everything

Further Comments on S

This should be right after 6.2 (p.56) (*)

1) S depends on the set of probabilities (p_α), not on the realisations of the random variable these probabilities ascribe (e.g. E_α). This means that change of variables ($E_\alpha \to f(E_\alpha)$) does not affect S (same uncertainty, same information)

2) Since $0 < p_\alpha < 1$, S is always non-negative

NB: If there is only one possible microstate, $p_\alpha = 1 \Rightarrow S = 0$ and vice versa because $-p_\alpha \ln p_\alpha \geq 0 \; \forall \alpha$.