03. Boltzmann Entropy, Gibbs Entropy, Shannon Information.

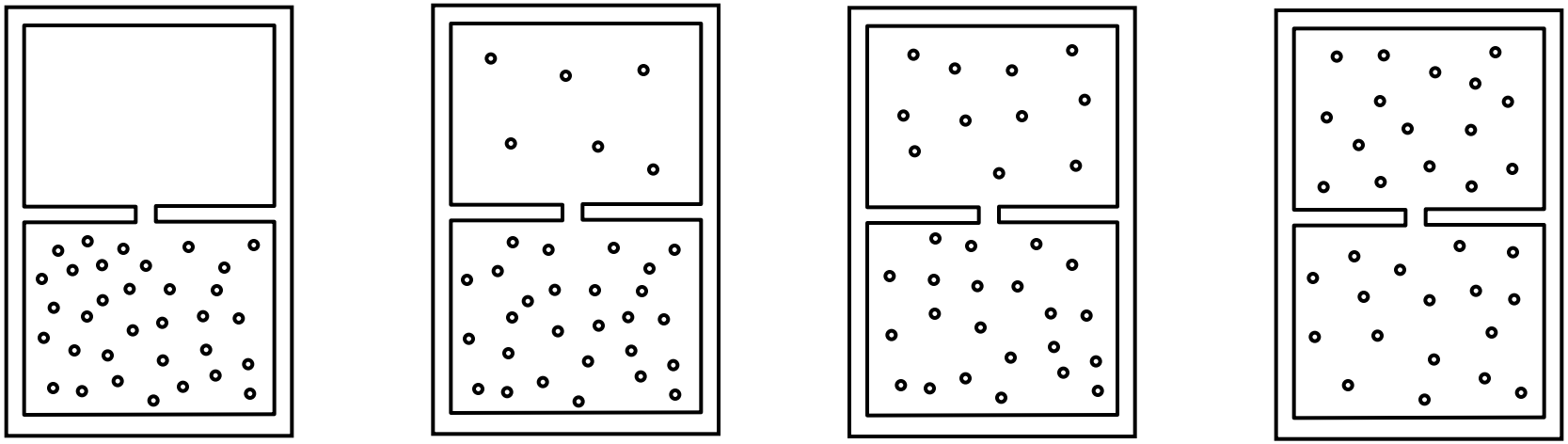
I. Entropy in Statistical Mechanics.

- *Goal*: To explain the behavior of *macroscopic* systems in terms of the dynamical laws governing their *microscopic* consituents.
  - *In particular*: To provide a *micro-dynamical* explanation of the 2nd Law.

Ludwig Boltzmann
(1844-1906)

1. Boltzmann's Approach.

- Consider different "macrostates" of a gas:



- Why does the gas prefer to be in the equilibrium macrostate (last one)?

Thermodynamic equilibrium macrostate = constant thermodynamic properties (temperature, volume, pressure, etc.)

- Suppose the gas consists of N *identical* particles governed by Hamilton's equations of motion (the micro-dynamics).

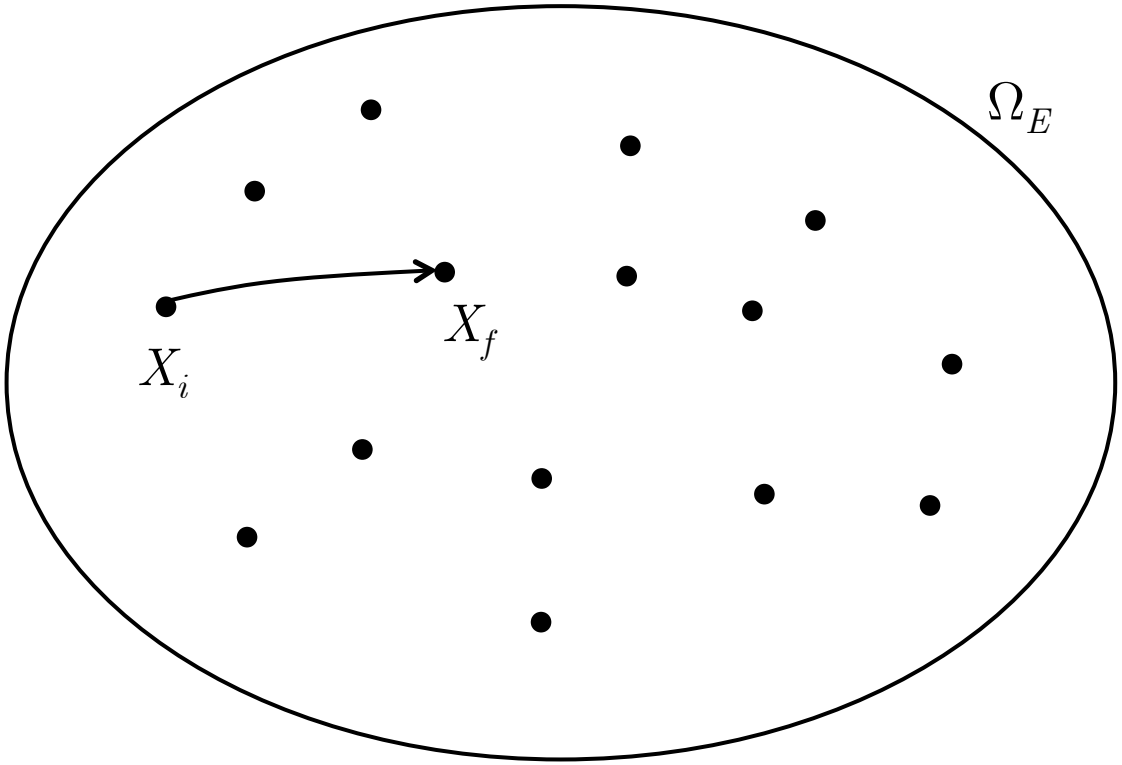
Def. 1. A *microstate* X of a gas is a specification of the position (3 values) and momentum (3 values) for each of its N particles.

Let Ω = *phase space* = $6N$-dim space of all possible microstates.

Let Ω_E = region of Ω that consists of all microstates with constant energy E.

Hamiltonian dynamics maps initial microstate X_i to final microstate X_f.

Can 2nd Law be explained by recourse to this dynamics?



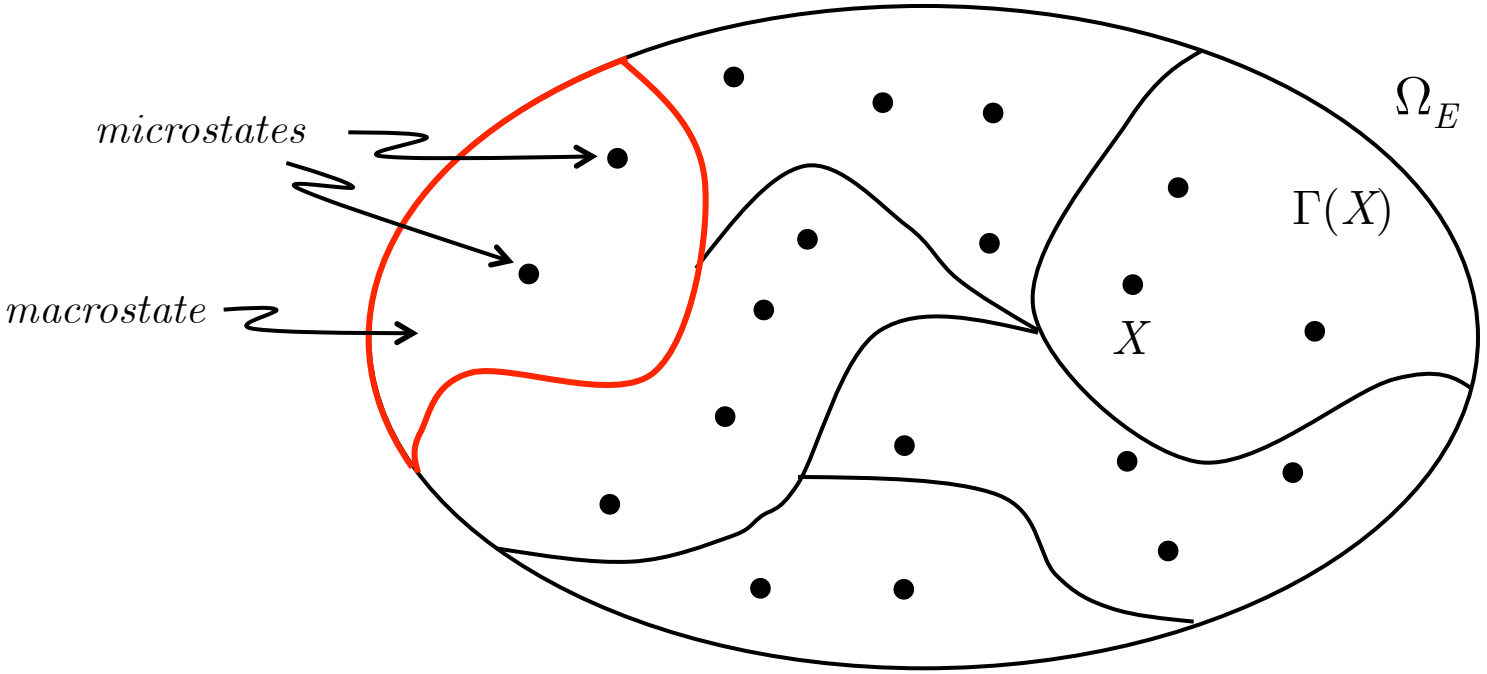
Def. 2. A *macrostate* Γ of a gas is a specification of the gas in terms of macroscopic properties (pressure, temperature, volume, *etc.*).

- *Relation between microstates and macrostates:*

Macrostates supervene on microstates!

  - *To each microstate there corresponds exactly one macrostate.*
  - *Many distinct microstates can correspond to the same macrostate.*

- *So*: Ω_E is partitioned into a finite number of regions corresponding to macrostates, with each microstate X belonging to one macrostate $\Gamma(X)$.



Boltzmann's Claim: The equilibrium macrostate Γ_{eq} is *vastly larger* than any other macrostate (so it contains the vast majority of possible microstates).

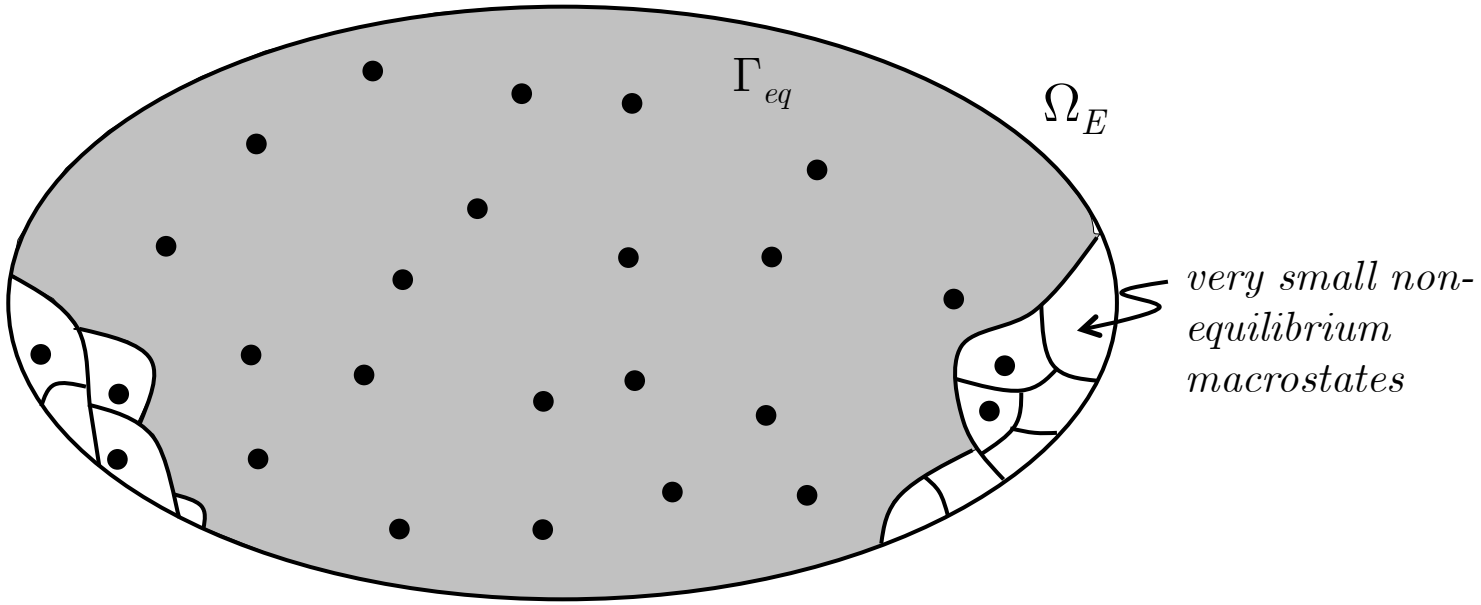
Def. 3. The *Boltzmann Entropy* is defined by

$$S_B(\Gamma(X)) = k \log|\Gamma(X)|$$

where $|\Gamma(X)|$ is the volume of $\Gamma(X)$.

So: $S_B(\Gamma(X))$ *is a measure of the size of* $\Gamma(X)$.

And: $S_B(\Gamma(X))$ obtains its maximum value for Γ_{eq}.



- *Thus*: S_B increases over time because, for any initial microstate X_i, the dynamics will map X_i into Γ_{eq} very quickly, and then keep it there for an extremely long time.

Two Ways to Explain the Approach to Equilibrium:

(a) Appeal to Typicality (Goldstein 2001)

> *Claim*: A system approaches equilibrium because equilibrium microstates are *typical* and nonequilibrium microstates are *atypical*.

- *Why?* For large N, Ω_E is almost entirely filled up with equilibrium microstates. Hence they are "typical".

- *But*: What is it about the *dynamics* that evolves atypical states to typical states?
  - "If a system is in an atypical microstate, it does not evolve into an equilibrium microstate *just because* the latter is typical." (Frigg 2009)
- Need to identify properties of the dynamics that guarantee atypical states evolve into typical states.
- *And*: Need to show that these properties are typical.
  - *Ex*: If the dynamics is *chaotic* (in an appropriate sense), then (under certain conditions), any initial microstate X_i will quickly be mapped into Γ_{eq} and remain there for long periods of time. (Frigg 2009)

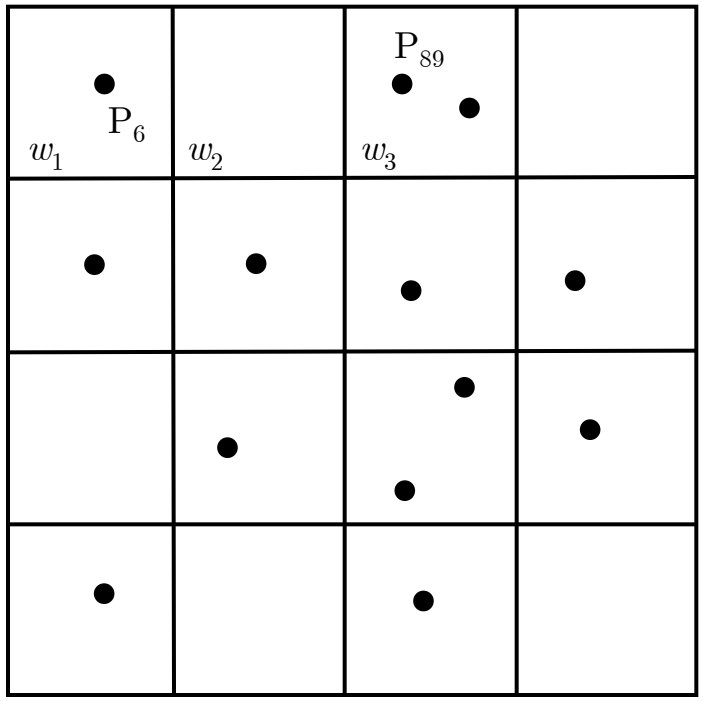
(b) Appeal to Probabilities

Claim: A system approaches equilibrium because it evolves from states of lower toward states of higher probability, and the equilibrium state is the state of highest probabililty.

- Associate probabilities with macrostates: the larger the macrostate, the greater the probability of finding a microstate in it.

"In most cases, the initial state will be a very unlikely state. From this state the system will steadily evolve towards more likely states until it has finally reached the most likely state, i.e., the state of thermal equilibrium."

Task: Make this a bit more precise (Boltzmann's combinatorial argument)...



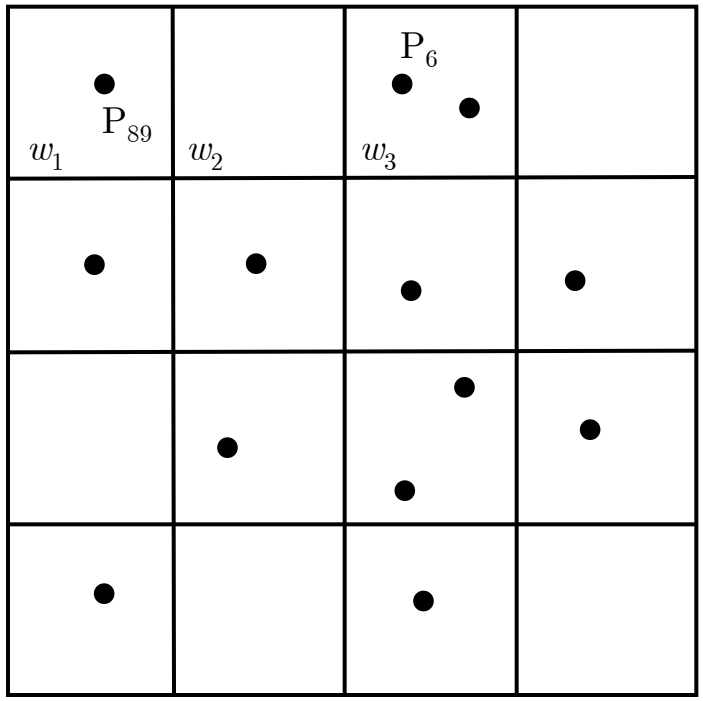
Arrangement #1:
state of P_6 in w_1, state of P_{89} in w_3, *etc.*

- Start with the *6-dim* phase space Ω_μ of a single particle.
- Partition Ω_μ into ℓ cells w_1, w_2, ..., w_ℓ of size δw.
- A state of an N-particle system is given by N points in Ω_μ.

$\Omega_E = N$ *copies of* Ω_μ

point in Ω_μ *= single-particle microstate.*

Def. 4. An *arrangement* is a specification of *which* points lie in which cells.



Arrangement #1:
state of P_6 in w_1, state of P_{89} in w_3, *etc.*

Arrangement #2:
state of P_{89} in w_1, state of P_6 in w_3, *etc.*

Distribution:
(1, 0, 2, 0, 1, 1, ...)

Takes form $(n_1, n_2, ..., n_\ell)$, where n_j = # of points in w_j.

- Start with the *6-dim* phase space Ω_μ of a single particle.
- Partition Ω_μ into ℓ cells w_1, w_2, ..., w_ℓ of size δw.
- A state of an N-particle system is given by N points in Ω_μ.

$\Omega_E = N$ *copies of* Ω_μ

point in Ω_μ = *single-particle microstate.*

Def. 4. An *arrangement* is a specification of *which* points lie in which cells.

Def. 5. A *distribution* is a specification of *how many* points (regardless of *which* ones) lie in each cell.

- *Note*: More than one arrangement can correspond to the same distribution.

- How many arrangements $G(D_i)$ are compatible with a given distribution $D_i = (n_1, n_2, ..., n_\ell)$?

- *Answer*: $$G(D_i) = \frac{N!}{n_1!\, n_2! \cdots n_\ell!}$$

$n! = n(n-1)(n-2)\cdots 1$
$= \#$ *of ways to arrange* n *distinguishable objects*
$0! = 1$

Number of ways to arrange N distinguishable objects into ℓ bins with capacities n_1, n_2, ..., n_ℓ.

Check: Let $D_1 = (N, 0, ..., 0)$ and $D_2 = (N-1, 1, 0, ..., 0)$.
- $G(D_1) = N!/N! = 1$. (Only one way for all N particles to be in w_1.)
- $G(D_2) = N!/(N-1)! = N(N-1)(N-2)\cdots 1/(N-1)(N-2)\cdots 1 = N$.
(There are N different ways w_2 could have one point in it; namely, if P_1 was in it, or if P_2 was in it, or if P_3 was in it, *etc*...)

"The probability of this distribution $[D_i]$ is then given by the number of permutations of which the elements of this distribution are capable, that is by the number $[G(D_i)]$. As the most probable distribution, i.e., as the one corresponding to thermal equilibrium, we again regard that distribution for which this expression is maximal..."

- *Again*: The *probability* of a distribution D_i is given by $G(D_i)$.

- *And*: Each distribution D_i corresponds to a macrostate Γ_{D_i}.

  Why? Because a system's macroscopic properties (volume, pressure, temp, *etc*) only depend on *how many* particles are in particular microstates, and not on *which* particles are in which microstates.

- *What is the size of this macrostate?*

  - A point in Ω_E corresponds to an arrangement of Ω_μ.
  - The size of a macrostate Γ_{D_i} in Ω_E is given by the number of points it contains (the number of arrangements compatible with D_i) multiplied by a *volume element* of Ω_E.
  - A volume element of Ω_E is given by N copies of a volume element δw of Ω_μ.

- *So*: The size of Γ_{D_i} is $|\Gamma_{D_i}| = \begin{bmatrix} \textit{number of} \\ \textit{arrangements} \\ \textit{compatible with } D_i \end{bmatrix} \times \begin{bmatrix} \textit{volume element} \\ \textit{of } \Omega_E \end{bmatrix}$

$$= G(D_i)\, \delta w^N$$

In other words: The probability $G(D_i)$ of a distribution D_i is proportional to the size of its corresponding macrostate Γ_{D_i}.

- *The equilibrium macrostate, being the largest, is the most probable; and a system evolves from states of low probabilty to states of high probability.*

- *And*: Each distribution D_i corresponds to a macrostate Γ_{D_i}.

  Why? Because a system's macroscopic properties (volume, pressure, temp, *etc*) only depend on *how many* particles are in particular microstates, and not on *which* particles are in which microstates.

- *What is the size of this macrostate?*

  - A point in Ω_E corresponds to an arrangement of Ω_μ.
  - The size of a macrostate Γ_{D_i} in Ω_E is given by the number of points it contains (the number of arrangements compatible with D_i) multiplied by a *volume element* of Ω_E.
  - A volume element of Ω_E is given by N copies of a volume element δw of Ω_μ.

- *So*: The size of Γ_{D_i} is $|\Gamma_{D_i}| = \begin{bmatrix} \textit{number of} \\ \textit{arrangements} \\ \textit{compatible with } D_i \end{bmatrix} \times \begin{bmatrix} \textit{volume element} \\ \textit{of } \Omega_E \end{bmatrix}$

$$= G(D_i)\, \delta w^N$$

- The Boltzmann entropy of Γ_{D_i} is given by:

$$\begin{aligned} S_B(\Gamma_{D_i}) &= k\log(G(D_i)\, \delta w^N) \\ &= k\log(G(D_i)) + Nk\log(\delta w) \\ &= k\log(G(D_i)) + \textit{const.} \end{aligned}$$

S_B is a measure of how large a macrostate is, and thus how probable the corresponding distribution of microstates is.

Other formulations of S_B

$$S_B(\Gamma_{D_i}) = k\log(G(D_i)) + const.$$

$$= k\log\left(\frac{N!}{n_1!n_2!\cdots n_\ell!}\right) + const.$$

$$= k\log(N!) - k\log(n_1!) - \ldots - k\log(n_\ell!) + const.$$

$$\approx (Nk\log N - N) - (n_1 k\log n_1 - n_1) - \ldots - (n_\ell k\log n_\ell - n_\ell) + const.$$

$$= \boxed{-k\sum_{j=1}^{\ell} n_j \log n_j + const.}$$

Stirling's approx:
$\log n! \approx n\log n - n$

$n_1 + \ldots + n_\ell = N$

S_B in terms of microstate occupation numbers n_j.

- *Let*: $p_j = n_j/N =$ (*probability of finding a randomly chosen microstate in cell w_j*)

Probabilities for *microstates*, not macrostates/distributions!

- *Then*: $\boxed{S_B(\Gamma_{D_i}) = -Nk\sum_{j=1}^{\ell} p_j \log p_j + const.}$

S_B in terms of microstate probabilities p_j.

The biggest value of S_B is for the distribution D_i for which the p_j's are all $1/\ell$; *i.e.*, the n_j's are all N/ℓ (the *equilibrium distribution*).

Relation between Boltzmann Entropy S_B and Thermodynamic Entropy S_T

- *First*: Let's derive the Maxwell-Boltzmann equilibrium distribution.
- *Assume*: A system of N *weakly interacting* particles described by:

$$\sum_j n_j = N$$

n_j = # *states in cell* w_j
N = *total # particles*

$$\sum_j \varepsilon_j n_j = U$$

ε_j = *energy of microstate in* w_j
U = *total internal energy*

Weakly ineracting assumption means total internal energy is just sum of energies of each particle

- *Recall*: $S_B(n_j) = (Nk\log N - N) - k\sum_j (n_j \log n_j - n_j) + const.$
- *So*:

$$\begin{aligned}\frac{d}{dn_j} S_B &= 0 - k\sum_j \frac{d}{dn_j}(n_j \log n_j - n_j) + 0 \\ &= -k\sum_j (\log n_j + n_j/n_j - 1) \\ &= -k\sum_j \log n_j\end{aligned}$$

- *Or*: $dS_B = -k\sum_j \log n_j \, dn_j$

- *Now*: S_B takes its maximum value for the values n_j^* that solve:

$$dS_B = -k\sum_j \log n_j^* \, dn_j = 0$$

Small changes to S_B due only to small changes dn_j.

subject to the constraints on the small changes dn_j:

$$dN = \sum_j dn_j = 0 \qquad dU = \sum_j \varepsilon_j \, dn_j = 0$$

- *Note*: Can add arbitrary multiples of the constraints to our equation and still get zero result:

$$dS_B = \sum_j (-k\log n_j^* + \alpha + \beta\varepsilon_j)\, dn_j = 0$$

- *Or*: $-k\log n_j^* + \alpha + \beta\varepsilon_j = 0$

- *Now solve for n_j^**:

$$n_j^* = e^{(\alpha + \beta\varepsilon_j)/k}$$

Maxwell-Boltzmann equilibrium distribution for weakly interacting, distinguishable particles. (Independently derived by Maxwell in 1860.)

What is α?

- Substitute n_j^* into $\sum n_j = N$ and get: $N = \sum e^{(\alpha + \beta\varepsilon j)/k} = e^{\alpha/k}\sum e^{\beta\varepsilon j/k}$.

- *Or*: $\alpha = k\log(N/\sum e^{\beta\varepsilon j/k})$ — *α is a normalization constant that enforces correct particle number N*

More importantly: What is β?

- *Consider*: Small changes in internal energy of a reversible process:

Macroscopic point of view	*Microscopic point of view*
$dU = \delta Q - dW$	$dU = d(\sum \varepsilon_j n_j)$
$= TdS_T - PdV$	$= \sum \varepsilon_j dn_j + \sum n_j d\varepsilon_j$

Two ways U can change:
- *ε_i changes, n_i constant (work)*
- *n_i changes, ε_i constant (heat)*

- *Note*: *If* $PdV = -\sum n_j d\varepsilon_j$, *then* $dS_T = (1/T)\sum \varepsilon_j dn_j$

- *Recall*: $dS_B(n_j^*) = -k\sum \log n_j^* dn_j$

$= -k\sum[(\alpha + \beta\varepsilon_j)/k]dn_j$, *since* $\log(e^{(\alpha+\beta\varepsilon_j)/k}) = (\alpha+\beta\varepsilon_j)/k$

$= -\beta\sum \varepsilon_j dn_j$, *since* $-k\alpha\sum dn_j = 0$

- *So*: *For the equilibrium distribution* n_j^*, $S_B = S_T$ *provided* $\beta = -1/T$.

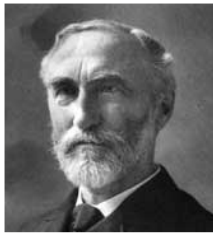
Macroscopic point of view	*Microscopic point of view*
$dU = \delta Q - dW$	$dU = d(\sum \varepsilon_j n_j)$
$= TdS_T - PdV$	$= \sum \varepsilon_j dn_j + \sum n_j d\varepsilon_j$

- *Recap*: $dS_B(n_j^*) = -\beta \sum \varepsilon_j dn_j$
- *So*: *For the equilibrium distribution* n_j^*, $S_B = S_T$ *provided* $\beta = -1/T$.

- *What this shows*: For a *reversible* process involving a *large number* of *distinguishable* particles *characterized by their positions and velocities*, it is *consistent* to identify the Boltzmann entropy S_B with the thermodynamic entropy S_T.
- *But*: Are we forced to?

- *S_T measures absolute changes in heat per temperature of a reverisble process.*	- *S_B measures how likely a given distribution of states occurs.*
- *For thermally isolated processes, S_T absolutely increases or remains constant.*	- *No absolute law that requires S_B to increase or remain constant.*

2. Gibbs' Approach.



Willard Gibbs
(1839-1903)

Boltzmann: Analysis of a *single* multiparticle system.

- Point x in Ω = possible microstate of system.
  - Thermodynamic property = function f on Ω.
- *Boltzmann equilibrium* macrostate = largest macrostate in Ω.

Thermodynamic equilibrium state = constant thermodynamic properties (temperature, volume, pressure, etc.)

Gibbs: Analysis of an *ensemble* of infinitely many copies of same system.

- Point x in Ω = actual state of one member of ensemble.
- State of entire ensemble = *distribution* $\rho(x,t)$ on Ω. ⟵ *Not Boltzmann's D!*
  - $\int_S \rho(x,t)\,dx$ = probability of finding the state of a system in region S.
  - Ensemble average of $f = \langle f \rangle = \int_\Omega f(x)\rho(x,t)\,dx$
- *Statistical equilibrium* distribution = stationary ρ (constant in time).
  - $\langle f \rangle$ is constant just when ρ is stationary.

So: If thermodynamic properties are represented by ensemble averages, then they don't change in time for an ensemble in statistical equilibrium.

Averaging Principle: The measured value of a thermodynamic property f of a system in thermodynamic equilibrium is the ensemble average $\langle f \rangle$ of an ensemble in statistical equilibrium.

Justification: A measurement of a property f takes some amount of time, which is "infinite" compared to molecular processes.

- *So*: What gets measured in the lab is the infinite time average $f^*(x_0)$:

$$f^*(x_0) = \lim_{\tau \to \infty} \frac{1}{\tau} \int_{t_0}^{t_0+\tau} f(\phi_t(x_0))\, dt$$

- *And*: For "ergodic" systems, $\langle f \rangle = f^*(x_0)$.

- The **Gibbs Entropy**: $S_G(\rho) = -k \int_\Omega \rho(x,t) \log(\rho(x,t))\, dx$

The ensemble average of the function $-k\log(\rho(x,t))$.

- How to choose an appropriate distribution ρ:
  - *Require it be stationary (statistical equilibrium).*
  - *Require that* $S_G(\rho)$ *be maximal.*

Interpretive Issues:

(1) *Why* do low-probability states evolve into high-probability states? (What justifies a given stationary, S_G-maximizing distribution $\rho(x, t)$?)

- *Characterizations of the dynamics are, again, required to justify this.*

(2) *How* are the probabilities to be interpreted?

(a) *Ontic probabilities* = properties of physical systems

- *Long run frequencies?*
- *Single-case propensities?*

(b) *Epistemic probabilities* = measures of degrees of belief

- *Objective (rational) degrees of belief?*
- *Subjective degrees of belief?*

II. Entropy in Classical Information Theory.

- *Goal*: To construct a measure for the amount of information associated with a message.

The amount of info gained from the reception of a message depends on how *likely* it is.

Claude Shannon
(1916-2001)

- The less likely a message is, the more info gained upon its reception!
- Let $X = \{x_1, x_2, ..., x_\ell\}$ = set of ℓ messages.

Def. 1. A *probability distribution* $P = (p_1, p_2, ..., p_\ell)$ on X is an assignment of a probability $p_j = p(x_j)$ to each message x_j.

- *Recall*: This means $p_j \geq 0$ and $p_1 + p_2 + ... + p_\ell = 1$.

Def. 2. A *measure of information for* X is a real-valued function $H(X)$: $\{$*prob. distributions on* $X\} \to \mathbb{R}$, that satisfies:

- *Continuity.* $H(p_1, ..., p_\ell)$ is continuous.
- *Additivity.* $H(p_1q_1, ..., p_\ell q_\ell) = H(P) + H(Q)$, for probability distributions P, Q.
- *Monoticity.* Info increases with ℓ for uniform distributions: If $m > \ell$, then $H(Q) > H(P)$, for any $P = (1/\ell, ..., 1/\ell)$ and $Q = (1/m, ..., 1/m)$.
- *Branching.* $H(p_1, ..., p_\ell)$ is independent of how the process is divided into parts.
- *Bit normalization.* The average info gained for two equally likely messages is one bit: $H(\frac{1}{2}, \frac{1}{2}) = 1$.

Claim (Shannon 1949): There is exactly one function that satisfies these criteria; namely, the *Shannon Entropy* (or *Shannon Information*):

$$H(X) = -\sum_{j=1}^{\ell} p_j \log_2 p_j$$

- $H(X)$ is maximal for $p_1 = p_2 = ... = p_\ell = 1/\ell$.
- $H(X) = 0$ just when one p_j is 1 and the rest are 0.
- Logarithm is to base 2: $\log_2 x = y \Rightarrow x = 2^y$.

Bit normalization requires:
- If $X = \{x_1, x_2\}$, and $P = (\frac{1}{2}, \frac{1}{2})$, then $H(X) = 1$.
- *Note*: $H(X) = -(\frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2}) = \log 2$.
- *And*: $\log 2 = 1$ *if and only if* log is to base 2.

In what sense is this a measure of information?

1. $H(X)$ as Maximum Amount of Message Compression

- Let $X = \{x_1, ..., x_\ell\}$ be a set of letters from which we construct the messages.
- Suppose the messages have N letters a piece.
- The probability distribution $P = (p_1, ..., p_\ell)$ is now over the letter set.

> *What this means*:
> - Each letter x_i has a probability of p_i of occuring in a message.
> - *In other words*: A typical message will contain $p_1 N$ occurrences of x_1, $p_2 N$ occurrences of x_2, *etc.*

- *Thus*:

$$\left(\begin{array}{l} \textit{The number of distinct} \\ \textit{typical messages} \end{array} \right) = \frac{N!}{(p_1 N)!(p_2 N)! \cdots (p_\ell N)!}$$

Number of ways to arrange N distinct letters into ℓ bins with capacities $p_1 N$, $p_2 N$, ..., $p_\ell N$.

- *So*:

$$\log_2 \left(\begin{array}{l} \textit{The number of distinct} \\ \textit{typical messages} \end{array} \right) = \log_2 \left(\frac{N!}{(p_1 N)!(p_2 N)! \cdots (p_\ell N)!} \right)$$

Let's simplify the RHS...

$$\log_2\left(\frac{N!}{(p_1N)!(p_2N)!\cdots(p_\ell N)!}\right) = \log_2(N!) - \{\log_2((p_1N)!) + ... + \log_2((p_\ell N)!)\}$$

$$\approx (N\log_2 N - N) - \{(p_1N\log_2 p_1N - p_1N) + ... + (p_\ell N\log_2 p_\ell N - p_\ell N)\}$$

$$= N\{\log_2 N - 1 - p_1\log_2 p_1 - p_1\log_2 N + p_1 - ... - p_\ell\log_2 p_\ell - p_\ell\log_2 N + p_\ell\}$$

$$= -N\sum_{j=1}^{\ell} p_j\log_2 p_j$$

$$= NH(X)$$

- <u>*Thus*</u>:

$$\log_2\left(\begin{array}{l}\textit{The number of distinct}\\ \textit{typical messages}\end{array}\right) = NH(X)$$

- <u>*So*</u>:

$$\left(\begin{array}{l}\textit{The number of distinct}\\ \textit{typical messages}\end{array}\right) = 2^{NH(X)}$$

- *So*: There are only $2^{NH(X)}$ typical messages with N letters.
- This means, *at the message level*, we can encode them using only $NH(X)$ bits.

> *Check*: 2 possible messages require 1 bit: 0, 1.
> 4 possible messages require 2 bits: 00, 01, 10, 11.
> *etc.*

- *Now*: *At the letter level*, how many bits are needed to encode a message of N letters drawn from an ℓ-letter alphabet?

> *First*: How many bits are needed to encode each letter in an ℓ-letter alphabet?
>
> | ℓ = #*letters* | x = #bits per letter | |
> |---|---|---|
> | 2 letters | 1 bit: | 0, 1 |
> | 4 letters | 2 bits: | 00, 01, 10, 11 |
> | 8 letters | 3 bits: | 000, 001, 010, 011, 100, 101, 110, 111 |
>
> *So*: $\ell = 2^x$, or $x = \log_2 \ell$

- *Note*: $\log_2 \ell$ bits per letter entails $N \log_2 \ell$ bits for a sequence of N letters.
- *Thus*: *If we know how probable each letter is*, instead of requiring $N \log_2 \ell$ bits to encode our messages, we can get by with only $NH(X)$ bits.
- *So*: $H(X)$ represents the *maximum amount that (typical) messages drawn from a given set of letters can be compressed.*

Ex: Let $X = \{A,\ B,\ C,\ D\}$ $\quad (\ell = 4)$

For instance:
$A = 00$, $B = 01$, $C = 10$, $D = 11$.

- *Then*: We need $\log_2 4 = 2$ bits per letter.
- *So*: We need $2N$ bits to encode a message with N letters.
- *Now*: Suppose the probabilities for each letter to occur in a typical N-letter message are the following:

$$p_A = 1/2, \quad p_B = 1/4, \quad p_C = p_D = 1/8$$

- *Then*: The minimum number of bits needed to encode all possible N-letter messages is:

$$NH(X) = -N\left(\tfrac{1}{2}\log_2\tfrac{1}{2} + \tfrac{1}{4}\log_2\tfrac{1}{4} + \tfrac{1}{8}\log_2\tfrac{1}{8} + \tfrac{1}{8}\log_2\tfrac{1}{8}\right) = 1.75N$$

- *Thus*: If we know how probable each letter is, instead of requiring $2N$ bits to encode all possible messages, we can get by with only $1.75N$.
- *Note*: If all letters are equally likely (the equilibrium distribution), then $p_A = p_B = p_C = p_D = 1/4$.
- *And*: $NH(X) = -N\left(\tfrac{1}{4}\log_2\tfrac{1}{4} + \tfrac{1}{4}\log_2\tfrac{1}{4} + \tfrac{1}{4}\log_2\tfrac{1}{4} + \tfrac{1}{4}\log_2\tfrac{1}{4}\right) = 2N.$

How the message compression interpretation of H relates to S_B

Shannon

- N = # of letters in message.
- N-letter message.
- $\{x_1, ..., x_\ell\}$ = ℓ-letter alphabet.
- $(p_1, ..., p_\ell)$ = probability distribution over letters.
- p_j = probability that x_j occurs in a given message.
- Np_j = # of x_j's in typical message.

$$H(X) = -\sum_{j=1}^{\ell} p_j \log_2 p_j$$

- NH = minimum number of base 2 numerals ("bits") needed to encode a message composed of N letters drawn from set X.

Boltzmann

- N = # of single-particle microstates.
- N-microstate arrangement.
- $(n_1, ..., n_\ell)$ = ℓ-cell distribution.
- $(p_1, ..., p_\ell)$ = probabilty distribution over microstates.
- $p_j = n_j/N$ = prob that a w_j-microstate occurs in a given arrangement.
- Np_j = # of w_j-microstates in arrangement.

$$S_B(\Gamma_{D_i}) = -Nk\sum_{j=1}^{\ell} p_j \ln p_j + const.$$

- $S_B \sim NH$ = minimum number of base e numerals ("e-bits?") needed to encode an arrangement of N single-particle microstates.

2. $H(X)$ as a Measure of Uncertainty

- Suppose $P = (p_1, ..., p_\ell)$ is a probability distribution over a set of values $\{x_1, ..., x_\ell\}$ of a random variable X.

Def. 1. The *expected value* $E(X)$ of X is given by $E(X) = \sum_{j=1}^{\ell} p_j x_j$.

Def. 2. The *information gained* if X is measured to have the value x_j is given by $-\log_2 p_j$.

- *Motivation*: The greater p_j is, the more certain x_j is, and the less information should be associated with it.

- Then the expected value of $-\log_2 p_j$ is just the Shannon information:

$$E(-\log_2 p_j) = -\sum_{j=1}^{\ell} p_j \log_2 p_j = H(X)$$

- *What this means*:

  $H(X)$ tells us our expected *information gain* upon measuring X.

How does the uncertainty interpretation of H relates to S_B

Shannon

- X = random variable.
- $\{x_1, ..., x_\ell\} = \ell$ values.
- $(p_1, ..., p_\ell)$ = probabilty distribution over values.
- p_j = probability that X has value x_j upon measurement.
- $-\log_2 p_j$ = *information* gained upon measurement of X with outcome x_j.

$$H(X) = -\sum_{j=1}^{\ell} p_j \log_2 p_j$$

- $H(X)$ = expected information gain upon measurement of X.

Boltzmann

- X = single-particle microstate.
- $(n_1, ..., n_\ell) = \ell$-cell distribution.
- $(p_1, ..., p_\ell)$ = probabilty distribution over microstates.
- $p_j = n_j/N$ = probability that a microstate occurs in cell w_j.
- $-\ln p_j$ = *information* gained upon measurement of particle to be in microstate in cell w_j.

$$S_B(\Gamma_{D_i}) = -Nk\sum_{j=1}^{\ell} p_j \ln p_j + const.$$

- S_B/N = expected information gain upon determining the microstate of a particle.

Interpretive Issues:

(1) How should the probabilities $p(x_i)$ be interpreted?

- *Emphasis is on uncertainty*: The information content of a message x_i is a function of how uncertain it is, with respect to the receiver.
  - *So*: Perhaps the probabilities are *epistemic*.
  - *In particular*: $p(x_i)$ is a measure of the receiver's degree of belief in the accuracy of message x_i.
- *But*: The probabilities are set by the nature of the source.
  - If the source is not probabilistic, then $p(x_i)$ can be interpreted epistemically.
  - If the source is inherently probabilistic, then $p(x_i)$ can be interpreted as the *ontic* probability that the source produces message x_i.

2. How is Shannon Information/Entropy related to other notions of entropy?

Thermodynamic entropy:	$\Delta S = S_f - S_i = \int_{i,R}^{f} \frac{\delta Q_R}{T}$
Boltzmann entropy:	$S_B(\Gamma_{D_i}) = -Nk\sum_{j=1}^{\ell} p_j \ln p_j + const.$
Gibbs entropy:	$S_G(\rho) = -k\int_{\Omega} \rho(x,t)\ln(\rho(x,t))\,dx$
Shannon entropy:	$H(X) = -\sum_{j=1}^{\ell} p_j \log_2 p_j$

- Can statistical mechanics be given an information-theoretic foundation?
- Can the 2nd Law be given an information-theoretic foundation?