Russell Goyder Nov 2022

Physical Entropy and Information Entropy

Summary: An examination of the concept of entropy from both perspectives, in which we will see that they are really the same.

Outline

Sources

[M] Information Theory, Inference, and Learning Algorithms, David MacKay
[J] Probability Theory, The Logic of Science, Edwin Jaynes
[S] Statistical Mechanics, MIT Open Courseware, Matthew Schwartz
[C] Thermodynamics and Intro to Stat. Mech, Callen.
[Sh] A Mathematical Theory of Communication, Shannon 1948

1. Information Entropy

1.1 Log 1/p as information [M 4.1]

How much information is in the outcome of a "random" experiment? Consider a random variable x taking discrete values $a_0, a_1, \dots, a_{N-1}$. Let $\text{Prob}(x = a_i) \Rightarrow p_i$.

Info of this outcome is $h_i = \log_2 \frac{1}{p_i}$ in bits

base 2 - bits
e - nats
10 - digits?
26 - Letters?

additive for compound events

rare outcomes convey more info (if what happened is what usually happens, you didn't learn much)

Similar to 20 questions: "Game of 16"

I'm thinking of an integer from 0 to 15. How many yes/no questions are needed to find the number?

	$n=13$	$n=6$
(1) Is $n \geq 8$?	$13 > 8$, Yes	$6 < 8$, No
(2) Is $(n \bmod 8) \geq 4$?	$5 > 4$, Yes	$6 > 4$, Yes
(3) Is $(n \bmod 4) \geq 2$?	$1 < 2$, No	$6 > 2$, Yes
(4) Is $(n \bmod 2) > 0$ (ie $=1$)?	$1 > 0$, Yes	$0 \neq 1$, No.

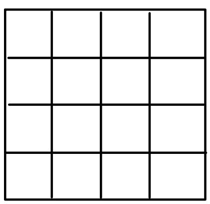
Answer is : 4 questions. If each n is equally probable, the outcome of each question has probability $p_i = 1/2$.

So $h_i = \log_2 2 = 1$ bit. Each question determines one bit in the binary representation of n:

$$13 \leftrightarrow 1101 \qquad 6 \leftrightarrow 0110.$$

What if the p_i are not all equal? Similar to "Battleships"

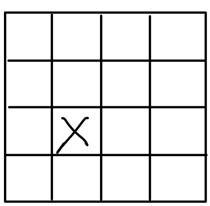
Game of Submarine



A submarine is hiding in one square.

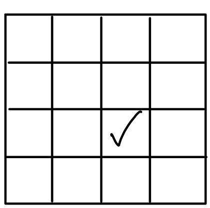
You guess squares one by one, I tell you hit (✓) or miss (X)

First turn:



Probability of a miss on the first turn

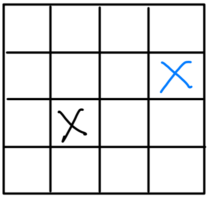
$$p_X^{(1)} = \frac{15}{16}.\quad h_X^{(1)} = \log_2 \frac{16}{15} \simeq 0.0931 \text{ bits}$$



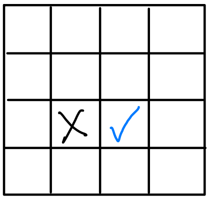
$$p_\checkmark^{(1)} = \frac{1}{16}.\quad h_\checkmark^{(1)} = \log_2 16 = 4 \text{ bits}$$

Probability of a hit on the first turn

2^{nd} turn:

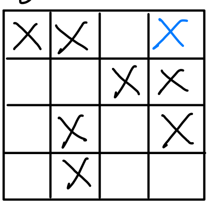


$$p_X^{(2)} = \frac{14}{15}, \quad h_X^{(2)} = \log_2 \frac{15}{14} \simeq 0.0995 \text{ bits}$$



$$p_{\checkmark}^{(2)} = \frac{1}{15}, \quad h_{\checkmark}^{(2)} = \log_2 \frac{15}{1} \simeq 3.91 \text{ bits}.$$

8^{th} turn:



$$\sum_{i=1}^{8} h_X^{(i)} = 1 \text{ bit}.$$ (We know it is not in half of the squares.)

12^{th} turn: $$\sum_{i=1}^{12} h_X^{(i)} = 2 \text{ bits}$$ (only 4 squares remaining)

14^{th} turn: $$\sum_{i=1}^{14} h_X^{(i)} = 3 \text{ bits}$$ (only 2 remaining)

15^{th} turn: $$\sum_{i=1}^{15} h_X^{(i)} = 4 \text{ bits}$$ (only 1 remaining)

While the cumulative information hits those round numbers of bits on certain special turns, the total info jumps to 4 bits whenever the submarine is hit!

$$\sum_{i=1}^{n-1} h_x^{(i)} + h_v^{(n)} = \log_2\left(\prod_{i=1}^{n-1} \frac{1}{p_i} \cdot \frac{1}{p_n}\right)$$

$$= \log_2\left(\frac{16}{15} \cdot \frac{15}{14} \cdot \cdots \cdot \frac{n+1}{n} \cdot \frac{n}{1}\right)$$

$$= \log_2 16 = 4 \text{ bits},$$ because the location of the submarine can be parametrized by an integer from 0 to 15.

1.2 Entropy as average information content

The average/expected amount of information in/learnable from a random variable('s probability distribution) is:

$$\langle h \rangle_p \equiv \sum_i p_i h_i = \underbrace{-\sum_i p_i \log_2 p_i}_{H_p \text{ (in bits)}} \text{ bits}$$

Entropy from Shannon's Axioms [Sh Appendix 2]

Information entropy $H_p \in \mathbb{R}$ is: $\longrightarrow$ generalize from rational to real probabilities

(1) Continuous

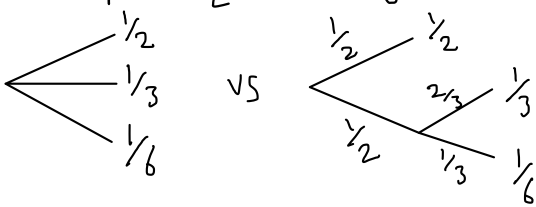
(2) Increasing $\longrightarrow$ If $p_i = \frac{1}{n}$ $\forall i \in \mathbb{N}^n$ then

$$A(n) \equiv H_p\Big(\underbrace{\tfrac{1}{n}, \tfrac{1}{n}, \ldots, \tfrac{1}{n}}_{n}\Big)$$

is an <u>increasing</u> function of n.

(3) Self-consistent.

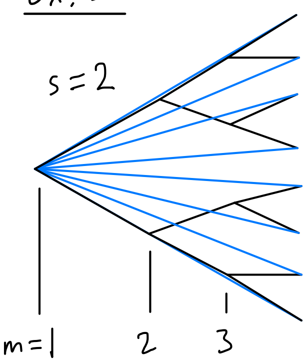
Example 1 [Sh I.6] (I: part, 6: section)



$$H\left(\tfrac{1}{2}, \tfrac{1}{3}, \tfrac{1}{6}\right) = H\left(\tfrac{1}{2}, \tfrac{1}{2}\right) + \frac{1}{2} H\left(\tfrac{2}{3}, \tfrac{1}{3}\right)$$

<u>Ex. 3</u>

$s = 2$

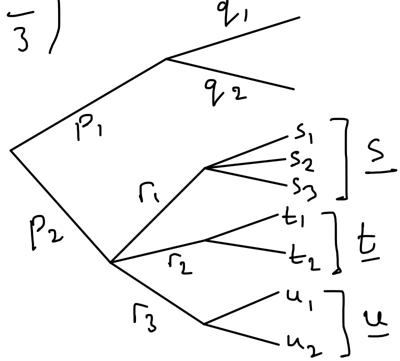


$s^m = 2^3 = 8$

$$A(s^m) = m\,A(s)$$

(see over)

<u>Ex. 2</u>

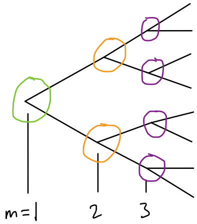


$$H(p_1q_1, p_1q_2, p_2r_1s_1, \ldots, p_2r_3u_2)$$

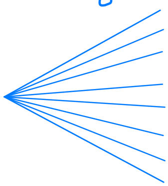
$$= H(p_1, p_2) + p_1 H(q_1, q_2)$$

$$+ p_2\left[H(r_1, r_2, r_3) + r_1 H(\underline{s}) + r_2 H(\underline{t}) + r_3 H(\underline{u})\right]$$

$A(s^m) = mA(s)$ for $s=2,\ m=3$



$$H\big(\underbrace{\tfrac{1}{8},\dots,\tfrac{1}{8}}_{8}\big) \equiv A(8) = A(2) + \frac{1}{2}\left[A(2) + \frac{1}{2}A(2) + \frac{1}{2}A(2)\right] + \frac{1}{2}\left[A(2) + \frac{1}{2}A(2) + \frac{1}{2}A(2)\right]$$



$$= A(2) + \frac{1}{2}\left[2A(2)\right] + \frac{1}{2}\left[2A(2)\right]$$

$$= \underline{3A(2)}$$

here, $m=2$

Let $p_i = \dfrac{n_i}{N}$ $\quad N = \sum_{i=1}^{n} n_i \quad n_i \in \mathbb{N}$

$$A(N) = H(p_1, p_2, \dots, p_n) + \sum_{i=1}^{n} p_i \underbrace{A(n_i)}$$

n_i equal possibilities with probability p_i

Shannon shows that:

$A(k) = C\log k$ for $C \in \mathbb{R}$

$k \in \mathbb{N} > 0$, using axioms (2) and (3)

$$C\log \sum n_i = H_p + C\sum_i p_i \log n_i$$

$$\therefore H_p = -C\sum_{i=1}^{n} p_i \log \underbrace{\frac{n_i}{N}}_{p_i} \quad \text{using} \quad \sum_{i=1}^{n} p_i = 1$$

absorb into choice of base for logarithm.

2. Physical Entropy

The work of Carnot, Clausius, Joule and others in the 1820's – 40's resulted in the idea of entropy S as:

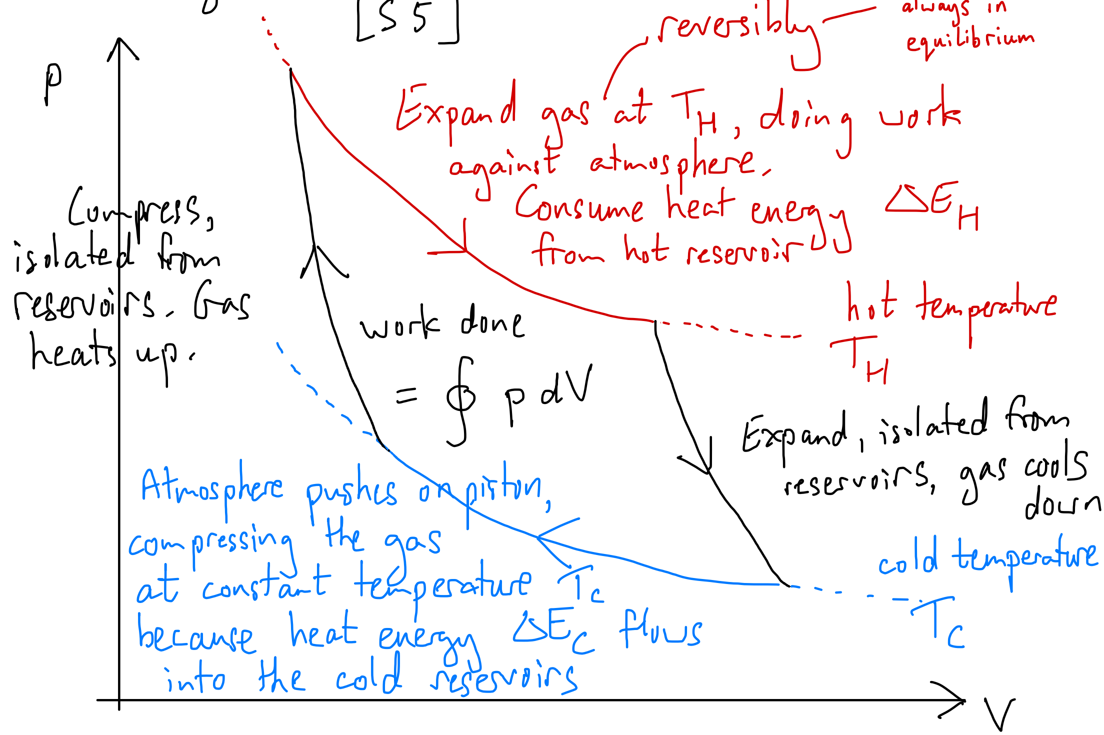
$$\text{temperature} \longrightarrow \frac{1}{T} = \left.\frac{\partial S}{\partial E}\right|_{V, N}$$

V — volume of gas

N — number of gas particles

E — internal energy of gas (ignore gravity etc.)

Carnot Cycle [C 4.7] [S 5]



Clausius found that over the cycle,

$$\oint \frac{dE}{T} = 0$$

Ie some quantity tracks the state of the system along the path:

$$dS = \frac{dE}{T}$$

2.1 Microstates and Macrostates [S 4.2]

Lecture ↓ Section

A microstate is a specific state of a physical system at the finest possible detail, such as the positions and momenta of all of the particles in an ideal gas.

A macrostate is a tuple of variables you care about and can measure, such as internal energy E, volume V and number of particles N of an ideal gas.

indirectly

Define $\Omega(E, V, N)$ as the number of microstates compatible with the macrostate E, V, N. If the gas is a mixture of two gases with energies E_1 and $E_2 = E - E_1$, then the number of microstates is

$$\Omega(E, E_1) = \Omega_1(E_1)\,\Omega_2(E - E_1)$$

\# microstates of gas 1
2

Probability of finding gas mixture in a state where gas 1 has energy E_1 is

$$P(E_1) = \underset{\text{const}}{C} \cdot \Omega_1(E_1)\,\Omega_2(E-E_1)$$

(each microstate ie equally likely by postulate)

Most probable value of E_1 satisfies $\left.\frac{\partial P(E_1)}{\partial E_1}\right|_E = 0$

$$0 = C\left(\frac{\partial \Omega_1}{\partial E_1}\cdot\Omega_2 - \Omega_1\frac{\partial \Omega_2}{\partial E_2}\right) \quad \text{because } \frac{\partial}{\partial E_2} = -\frac{\partial}{\partial E_1}$$

$$\Rightarrow \left.\frac{\partial \log \Omega_1(E)}{\partial E}\right|_{E=E_1} = \left.\frac{\partial}{\partial E}\log\Omega_2(E)\right|_{E=E_2}$$

So, there is a quantity $\beta \equiv \frac{\partial \log \Omega(E)}{\partial E}$ which is equal in equilibrium:

$$\beta_1 = \beta_2$$

Boltzman's Entropy

With $\beta = \frac{1}{k_B T}$ and $S = k_B \log \Omega$,

we recover $\frac{1}{T} = \left.\frac{\partial S}{\partial E}\right|_{N,V}$

Boltzman's entropy links thermodynamics to statistical mechanics by inviting us to count microstates via Ω.

2.2 Counting Microstates [J 11.4]

The macrostate is what we know and the microstate is unknown. What is knowable in practice? Measurement is subject to a level of accuracy or resolution, so the N particles in our ideal gas each fall into one of m groups, corresponding to regions of phase space.

So, $\Omega = \dfrac{N!}{n_1! n_2! \dots n_m!}$ where n_i particles are in the i'th group.

Stirling: $\log n! = n \log n - n + O(\log n)$

$$\log \Omega \simeq N \log N - \cancel{N} - \sum_{i=1}^{m} n_i \log n_i + \cancel{\underbrace{\sum_{i=1}^{m} n_i}_{N}}$$

$$= N\left(-\sum_i p_i \log n_i + \log N \underbrace{\sum_i p_i}_{1}\right) \qquad p_i = \frac{n_i}{N}$$

$$= -N\left(\sum_i p_i \left(\log n_i - \log N\right)\right)$$

$$\therefore \quad S = -k_B N \sum_{i=1}^{m} p_i \log p_i$$

If we have full knowledge of the system, $\Omega = 1$ and $S = 0$.

Physical entropy $S = -k_B N \sum_{i=1}^{m} p_i \log p_i$ and information entropy $H_p = -\sum_{i=1}^{m} p_i \log_2 p_i$ are therefore related by $S = \frac{k_B N}{\log_2 e} H_p$, where $\frac{k_B N}{\log_2 e}$ just changes units.

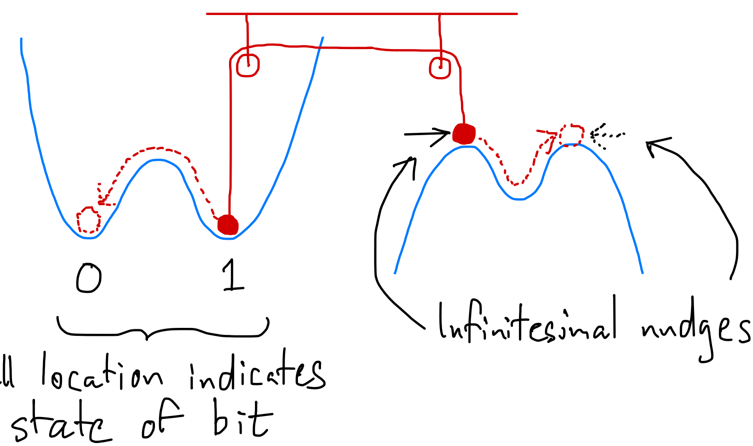
They are essentially the same thing. The work of Landauer (1961) and Bennet (1982) makes the connection stronger.

3. Connections

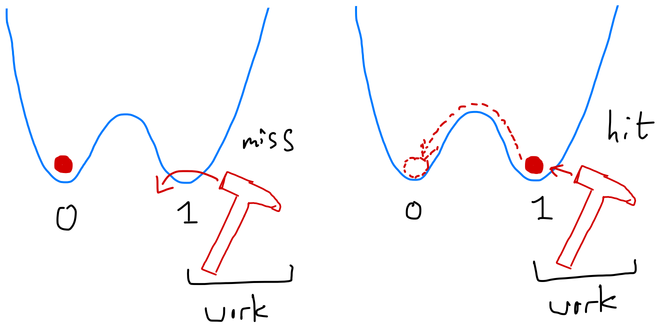
3.1 Landauer's Principle [S 6.5.2]

Erasing information:
costs energy / dissipates heat / increases entropy (elsewhere)

Just flipping a bit doesn't cost energy:



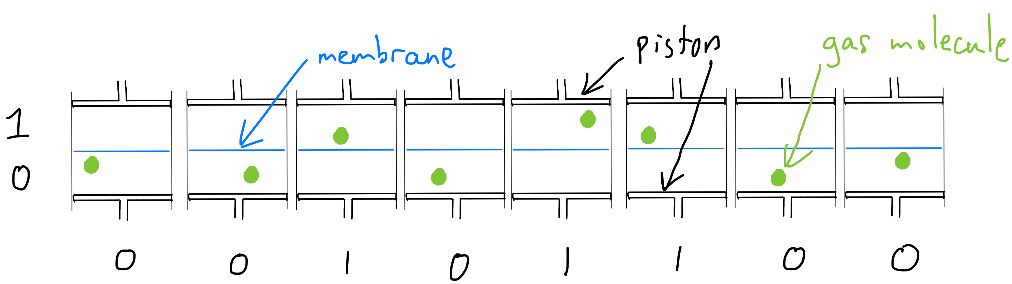
But ensuring the bit is set to 0 regardless of where it starts costs energy, because all known physical laws are time-reversible, yet erasure is not.



Even hitting balls with hammers involves time-reversible laws of physics, but by using the hammer we dissipate heat into parts of the system (computer!) that we are not using to store information.

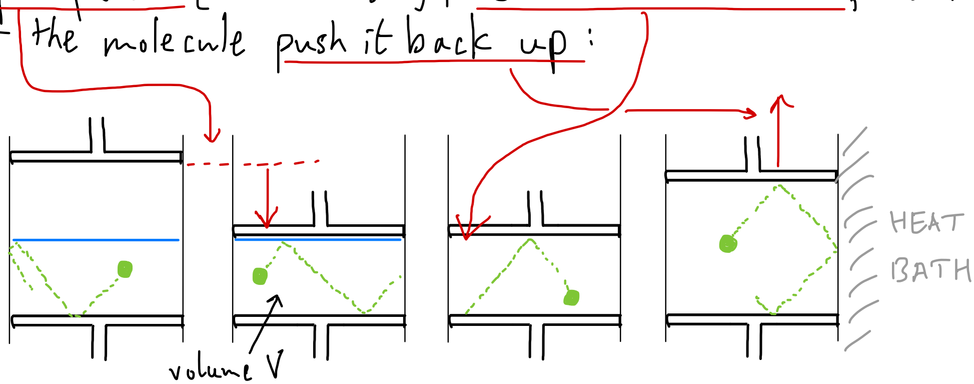
3.2 Szilard's Engine [S 6.5.3]

Bennet (1982) considered a "molecular tape" information storage device



Bits are encoded according to whether the molecule is in the lower or upper half of the cylinder. The temperature is kept constant, so the speed of each molecule is the same and the only variable of interest is its binary location.

If we know a molecule is in the lower half, can compress upper piston (at no cost), remove the membrane, then let the molecule push it back up:



This is Szilard's Engine. Energy extracted from the heat bath in one cycle for an ideal gas obeying $pV = k_B T$ $(N=1)$:

$$W = \int_V^{2V} p\, dV = k_B T \int_V^{2V} \frac{dV}{V} = k_B T \log 2$$

Afterwards, we have lost 1-bit of information because the molecule is free to move through the entire cylinder.

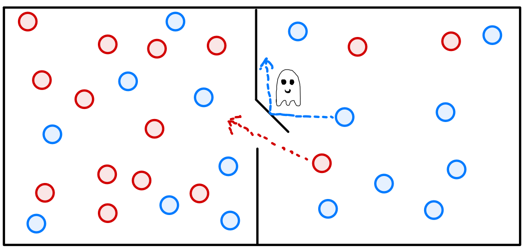
The (information in the) tape has a "fuel value" which must be included in any analysis of a system that uses it.

3.3 The 2nd Law and Maxwell's Daemon [S 6.6]

Challenging the 2nd law of thermodynamics : $\Delta S \geq 0$.

"The entropy of the universe tends to increase".

Posed in 1867, Szilard was working on it in the 1920s.



high pressure

low pressure

The daemon lets red particles through right to left, but not blue.

Red and blue are arbitrary labels, eg:

Red	Blue
Helium	Xenon
Fast-moving	Slow-moving
On the right	On the left

Could be all the same type of particle, just with different speeds or even positions

If the daemon can do this without being affected itself, then it violates the 2nd law, because isolating the red particles on the left hand side reduces their entropy.

Many attempts to resolve this over >100 years, eg consciousness, energy of photon required to observe particle... but daemon can be a machine / mechanism, energy of photon can be made negligibly small.

Ultimately, the daemon must acquire the knowledge of red vs blue, and remember it for at least a little while. This requires information storage, which in turn requires a bit to set to zero, which requires work and sends $k_B T \log 2$ Joules of heat into the universe, which exactly offsets the entropy decrease from isolating a red particle on the left hand side of the box.

Appendix: Maximum Entropy Distribution(s) [S 4.6]

If group i has energy ε_i, then $E = \sum_{i=1}^{m} n_i \varepsilon_i$ and the average energy $\bar{\varepsilon} = \frac{E}{N} = \sum_i p_i \varepsilon_i$.

Maximize:

$$L(p_i, \alpha, \beta) = -\sum_{i=1}^{m} p_i \log p_i - \alpha\underbrace{\Big(\sum_i p_i - 1\Big)}_{\text{conserve particles}} - \beta\underbrace{\Big(\sum_i p_i \varepsilon_i - \bar{\varepsilon}\Big)}_{\text{fix average energy}}$$

$$\frac{\partial L}{\partial p_j} = -\log p_j - 1 - \alpha - \beta\varepsilon_j = 0 \Rightarrow p_i = e^{-1-\alpha} e^{-\beta\varepsilon_i}$$

Exponential distribution given known $\bar{\varepsilon}$.
If know second moment, get quadratic term $\rightarrow$ Gaussian.

$$\sum_i p_i = e^{-1-\alpha} \underbrace{\sum_i e^{-\beta\varepsilon_i}}_{Z} = 1 \Rightarrow e^{1+\alpha} = Z$$

$$\therefore \quad p_i = \frac{e^{-\beta\varepsilon_i}}{Z}$$

Using $\bar{\varepsilon} = NE$ in L and that $L = \frac{S}{k_B N}$

$$\frac{1}{T} = \left.\frac{\partial S}{\partial E}\right|_N = k_B \cancel{N} \cdot \frac{\beta}{\cancel{N}} \Rightarrow \beta = \frac{1}{k_B T}.$$

So, $p_i = \dfrac{e^{-\frac{\varepsilon_i}{k_B T}}}{Z}$ where $Z = \sum_{i=1}^{m} e^{-\frac{\varepsilon_i}{k_B T}}$ Partition function

If remove the β constraint, get $p_i = \frac{1}{m}$ (each group equally likely)

Boltzman Distribution