1 Introduction to Information and Entropy

Let $X$ be a random variable with discrete outcomes, $\{x_1, \ldots, x_M\}$, where $M \geq 2$. Examples are as follows:

1. A coin having two outcomes of Head and Tail.
2. A six-faced die having six outcomes of 1, 2, 3, 4, 5, and 6.
3. An ideal gas system, contained in a rigid, impermeable (i.e., non-porous), and diathermal vessel, having with $N$ molecules. Under (quasi-static) thermodynamic equilibrium conditions, the system may have $M$ energy states, where $2 \leq M \ll N$.

Let $X$ have a probability distribution, described by a probability mass function $f_{x_i}$, where $\sum_{i=1}^{M} p_i = 1$ and $p_i \geq 0 \forall i$. Now, we pose the following question:

What is the information content of $X$ (i.e., the probability mass function $f_{x_i}$)?

To answer the above question, let us construct a message of finite length $N$, where $N \gg M$. The message could be constructed from independent realizations of the random variable $X$. Let us find out how many binary digits (called bits) are needed to convey this $N$-long message. If there are $R$ bits, then it follows that

$$2^R = M^N \Rightarrow R \log 2 = N \log M \Rightarrow R = \frac{N \log M}{\log 2}$$

Taking the the logarithm with base 2, it follows that $R = N \log_2 M$. Apparently, an $N$-long string of independent outcomes of the random variable $X$ has an information content of $N \log_2 M$ bits, i.e., this many bits of information will have to be transmitted to convey the message.

The probability distribution $\{p_i\}$ limits the types of messages that are likely to occur. For example, if $p_j \gg p_k$, then it is very unlikely to construct a message with the number of $x_k$'s being larger than the number of $x_j$'s. For $N$ being very large, we expect that $x_j$ will appear approximately $Np_j$ times out of $N$. Therefore, a typical message will contain $\{n_i = Np_i; \ i = 1, \ldots, M\}$ symbols arranged in different ways. The number of different arrangements is given by

$$n(N, M) = \frac{N!}{n_1! \cdots n_M!} \text{ where } \sum_{i=1}^{M} n_i = N \text{ and } n_i \geq 0 \forall i$$

it is noted that $\eta \ll M^N$ which is the maximum possible number of $N$-long messages. Then, it follows by using Stirling formula (which states $\log_e k! = k \log_e k - k + O(\log_e k)$) that

$$\log_e \eta = \log_e N! - \sum_{j=1}^{M} \log_e n_j! \approx (N \log_e N - N) - \sum_{j=1}^{M} (n_j \log_e n_j - n_j)$$

$$= N \log_e N - \sum_{j=1}^{M} n_j \log_e n_j = N \log_e N - \sum_{j=1}^{M} (Np_j) \log_e (Np_j)$$

$$= N \log_e N - \left( N \log_e N \right) \left( \sum_{j=1}^{M} p_j \right) = N \sum_{j=1}^{M} p_j \log_e p_j$$

Then, to represent one of the "likely" $\eta$ sequences, it takes $\log_2 \eta \approx -N \sum_{j=1}^{M} p_j \log_2 p_j$ bits of information.
Shannon’s Theorem states that, as $N \to \infty$, the minimum number of bits necessary to ensure the errors to vanish in $N$ trials is $\log_2 \eta \approx -N \sum_{j=1}^{M} p_j \log_2 p_j$, which is less than $N \log_2 M$ bits as needed in the absence of any knowledge of the probability distribution $\{p_i\}$. The difference per trial can be attributed as the information content $I$ of the probability distribution $\{p_i\}$, i.e., $\frac{N \log_2 M + \sum_{i=1}^{M} p_i \log_2 p_i}{N}$, from which it follows that $I[\{p_i\}] \triangleq \log_2 M + \sum_{i=1}^{M} p_i \log_2 p_i$. An alternative representation of information (that is adopted by many authors) is: $I[\{p_i\}] \triangleq \sum_{i=1}^{M} p_i \log_2 p_i$.

2 A Thermodynamic Perspective

Let a vessel with rigid, impermeable, and diathermal boundaries contain $N$ (non-interacting and statistically independent) randomly moving particles. Under a thermodynamic equilibrium condition, let the total energy of these $N$ particles be $E$ that is distributed as follows:

Let $N$ particles be clustered in $M$ groups, where $M \geq 2$ and $M \ll N$. In group $i$, where $i = 1, 2, \cdots, M$, there are $n_i$ particles such that the expected value of the energy of each particle is $\varepsilon_i$ with standard deviation $\delta_i$. Then it follows that

$$N = \sum_{i=1}^{M} n_i \quad \text{and} \quad E = \sum_{i=1}^{M} n_i \varepsilon_i \quad \text{for a very large} \ N.$$  

Let us order these $M$ groups of particles such that $\varepsilon_1 < \varepsilon_2 < \cdots < \varepsilon_M$. It is assumed that $(\frac{\delta_i}{\varepsilon_i}) \ll 1$ and $(\frac{\sqrt{\varepsilon_i^2 + \delta_i^2}}{\varepsilon_{i+1} - \varepsilon_i}) \ll 1$. Let us define $p_i \overset{\triangleq}{=} \frac{n_i}{N}$ and $E_i \overset{\triangleq}{=} N \varepsilon_i$, where $i = 1, 2, \cdots, M$; obviously, $\sum_{i=1}^{M} p_i = 1$ and $E_1 < E_2 < \cdots < E_M$ and $E = \sum_{i=1}^{M} (p_i E_i)$. Then, it follows from the principle of energy minimization at an equilibrium condition that $p_1 > \cdots > p_M$.

Let us initiate a quasi-static change through exchange of energy via the diathermal boundaries so that the total energy of the thermodynamic system is now $\bar{E}$ while the number of particles is still the same. Then these $N$ particles have a new probability distribution $\bar{p}_i$, $i = 1, 2, \cdots, M$ because the $N$ particles are now distributed among the same groups as $\bar{n}_i$, $i = 1, 2, \cdots, M$. The following conditions hold at this new condition.

$$\sum_{i=1}^{M} \bar{p}_i = 1 \quad \text{and} \quad \bar{E} = \sum_{i=1}^{M} (\bar{p}_i E_i)$$

Note that $E_i$’s are unchanged and the expected value of the energy of each of the $\bar{n}_i$ particles in the $i^{th}$ group is still $\varepsilon_i$ for $i = 1, 2, \cdots, M$.

**Remark 2.1.** The particle energies $\varepsilon_i$ are discrete according to quantum mechanics and their values depend on the volume to which these particles are confined; therefore, the possible values of the total energy $E$ are also discrete. However, for a large volume and consequently a large number of particles, the spacings of the different energy values are so small in comparison to the total energy of the system that the parameter $E$ can be regarded as a continuous variable. Note that this fact prevails regardless of whether the particles are non-interacting or interacting.

**Remark 2.2.** For a general case, where the vessel boundaries are allowed to be flexible, porous, and diathermal, the specifications of the respective parameters $E$, $V$ and $N$ define a macrostate of the thermodynamic system. However, at the particle level, there is a very large number of ways in which a macrostate $(E, V, N)$ can be realized. As seen above in the case of non-interacting particles, the total energy is simply the sum of the energies of $N$ particles; since these $N$ particles can be arranged in many different ways, each single particle of energy $\varepsilon_i$ can be placed in many different ways to realize the total energy $E$. Each of these different ways specifies a macrostate of the system and the actual number $\Omega$ of these microstates is a function of $E$, $V$ and $N$. In general, the microstates of a given system are generated in quantum mechanics as the independent solutions in the form of wave functions $\psi(r_1, \cdots, r_N)$ of the Schrödinger equation corresponding to the eigenvalue $E$ of the relevant operator. In essence, a given macrostate of the system corresponds to a large number of microstates; in other words, a macrostate is an equivalence class of a large number of microstates. In the absence of any constraints, these microstates are equally probable, i.e., the system is equally likely to be in any one of these microstates at an instant of time.
Chapter 4: Information Theory - Part I

Information is viewed as a measure of the knowledge derived from observed data if the probability distribution (and nothing else) of the data is known.

**Definition 1:** Let $A$ be a (nonempty and finite) alphabet of symbols. Let $l$ be the length (i.e., # of symbols) in a pattern. Then, # of such patterns $N \leq |A|^l$ and a typical pattern is expressed as a word or string of $l$ symbols, concatenated together.

**Example:** If $A = \{0, 1\}$, i.e., $|A| = 2$, then $N \leq 2^l$, i.e., $\log_2 N \leq l$ and $\mathbf{w} = 0. b_1. b_2. \ldots. b_l$ where $b_i \in \{0, 1\}$, in binary notation.

**Definition 2:** Shannon information for a word of length $l$ on the alphabet $A$ is defined as:

$$I(\mathbf{p}) \equiv \mathbb{E} \left[ \log_2 \mathbf{p} \right] = \sum_{i=1}^{l} p_i \log_2(p_i)$$

where $\mathbf{p} = [p_1, \ldots, p_l]$ with $p_i \geq 0 \ \forall i \in \{1, \ldots, l\}$ and $\sum_{i=1}^{l} p_i = 1$.

**Remark:** Shannon entropy $S(\mathbf{p}) = -I(\mathbf{p})$. This concept is taken from equilibrium thermodynamics where the alphabet $A$ is equivalent to the collection of finite number of energy states and $p_i$ is the probability of $i$th energy state being occupied by a particle under (macroscopic) equilibrium.

**Remark:** If $p_j = 1$ for some $j \in \{1, \ldots, l\}$, then $p_i = 0 \ \forall i \neq j$.

In that case, $I(\mathbf{p}) = 0$, which is equivalent to $S(\mathbf{p}) = 0$.

If $\mathbf{p}$ is a uniform distribution, i.e., $p_i = \frac{1}{l} \ \forall i \in \{1, \ldots, l\}$.

In that case $I(\mathbf{p}) = -\log_2(l)$, which is equivalent to $S(\mathbf{p}) = \ln(l)$.
Khinchin Axioms (1957)

**Axiom #1:** $I(\mathcal{P}) = I(p_1, \ldots, p_k)$ only depends on $\mathcal{P}$

**Axiom #2:** $I\left(\frac{1}{k}, \ldots, \frac{1}{k}\right) \leq I(p_1, \ldots, p_k)$, i.e., minimum information under uniform distribution

**Axiom #3:** $I(p_1, \ldots, p_k) = I(p_1, \ldots, p_k; 0)$

Augmentation of the data set with a new symbol, whose probability is 0, does not change the information.

**Axiom #4:** $I(p_{\text{aug}}) = I(p^{\text{sys}}) + \sum_i p_i I(\mathcal{P} | i)$

where $I(\mathcal{P} | i) = \sum_j p_{ji} \log(\frac{p_{ji}}{p_{ij}})$ is the conditional information of the probability distribution of the added system $\mathcal{O}$. [Note: $p_{ji}$ is the probability of the event $j$ of the added system when event $i$ of the $\mathcal{O}$ system has occurred.]

$p_{\text{aug}}$ is the probability distribution of the augmented system $\mathcal{O}_{\text{aug}}$

$p^{\text{sys}}$ is the probability distribution of the original system $\mathcal{O}^{\text{sys}}$

In Axiom #4, the composition of the original system $\mathcal{O}^{\text{sys}}$ with the added system $\mathcal{O}$ yields the augmented system $\mathcal{O}_{\text{aug}}$. 
**Rényi Information**

Shannon information $I_p(p)$ is additive for independent subsystems. The fourth axiom of Rényi information for non-independent subsystems cannot be handled through usage of the Shannon information. Therefore, other concepts of information need to be pursued.

A new information, called Rényi information, satisfies Axioms I, II, III, and the additivity under subsystem independence but not the Axiom IV:

$$I_p^\beta(p) = \frac{1}{\beta-1} \log \left( \sum_{i=1}^{r} (p_i)^\beta \right) \quad \text{for } \beta \in \mathbb{R}$$

where $r$ is the # of non-empty states $i$, i.e., # of all events $i$ of the sample set with $p_i > 0$.

For $\beta > 0$, Rényi information $I_0^\beta(p) = -\log r$.

$$\Rightarrow |I_0^\beta(p)| \text{ grows logarithmically with the # of non-empty states}$$

Setting $\varepsilon \equiv \beta - 1$, as $\beta \to 1$, then

$$\sum_{i=1}^{r} (p_i)^{1+\varepsilon} = \sum_{i=1}^{r} p_i \varepsilon \log p_i \sim \sum_{i=1}^{r} p_i (1 + \varepsilon \log p_i)$$

$$= \sum_{i=1}^{r} p_i + \varepsilon \sum_{i=1}^{r} p_i \log p_i \approx 1 + \varepsilon I(p)$$

$$\Rightarrow \sum_{i=1}^{r} p_i^\beta \approx 1 + \varepsilon I(p)$$

Hence, \( \lim_{\varepsilon \to 0} I_{1+\varepsilon}(p) = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \log (1 + \varepsilon \sum_{i=1}^{r} p_i \log p_i) \)

$$= \sum_{i=1}^{r} p_i \log p_i = I(p)$$

$$\Rightarrow \lim_{\beta \to 1} I_p^\beta(p) = I(p)$$

Hence, $I^1(p) = I(p) \forall p$

Alternatively, one can use the L'Hôpital rule to obtain the above result.
Information Gain

Given a probability distribution \( \mathbf{p} \) with \( p_i > 0 \) \( \forall i \) over a sample set; let the \( \mathbf{p} \) be changed to another distribution \( \mathbf{p}' \). Then, the change in the bit number of the event \( i \) in the set, i.e.,

\[
b_i^0 - b_i = -\log p_i^0 - (-\log p_i) = \log \left( \frac{p_i}{p_i^0} \right)
\]

where \( b_i = -\log p_i \) and \( b_i = -\log p_i \) indicate missing knowledge of the observer w.r.t. the event \( i \). For example, if \( N \) is the total number of elementary events in the system and \( N_i \) is the number of elementary events \( i \Rightarrow p_i = \frac{N_i}{N} \)

We define \( b_i \) to satisfy the relation \( b_i + \log N_i = \log N \)

\[
\Rightarrow b_i = -\log p_i \quad \text{where} \quad b_i \text{ is the bit number of the missing knowledge.}
\]

**Definition**: Information gain or Kullback information from the nominal system \( \Theta^0 \) to the perturbed system \( \Theta \) is defined as the average change in the bit number:

\[
K(\mathbf{p}, \mathbf{p}^0) = \sum_i p_i^0 \log \left( \frac{b_i}{p_i^0} \right)
\]

where \( \Theta \) is characterized by \( \mathbf{p} = [p_1, p_2, \ldots, p_r] \)

\( \Theta^0 \) is characterized by \( \mathbf{p}^0 = [p_1^0, p_2^0, \ldots, p_r^0] \)

**Proposition 1**: \( K(\mathbf{p}, \mathbf{p}^0) > 0 \) and equality holds iff \( \mathbf{p} = \mathbf{p}^0 \)

**Proof**: The proof requires Jensen's inequality that is stated below.

**Jensen's Inequality**: If \( g: \mathbb{R} \to \mathbb{R} \) is a convex function, \( X \) is a random variable defined over the probability space, then \( E[g(X)] \geq g(E[X]) \). Furthermore, if \( g \) is strictly convex, then \( X = E[X] \) a.s.
**Definition:** A function $g: \mathbb{R} \to \mathbb{R}$ is called convex over an open interval $(a, b)$ if the following condition holds:

$$g(\lambda x_1 + (1-\lambda) x_2) \leq \lambda g(x_1) + (1-\lambda) g(x_2)$$

$$\forall x_1, x_2 \in (a, b) \quad \forall \lambda \in [0, 1].$$

Furthermore, the function $g$ is called strictly convex if the equality holds only for the boundary points $\lambda = 0$ and $\lambda = 1$.

**Proof of Jensen's inequality:**

The proof makes use of the method of induction. To start with, let us assume the sample space consists of two points, i.e., the distribution is $b$ and $1-b$. Then, convexity of $g$ implies:

$$b g(x_1) + (1-b) g(x_2) \geq g(b x_1 + (1-b) x_2)$$

Let Jensen's inequality be true for $n$ points, where $n \geq 2$.

Let $q_i = \frac{b_i}{1-b_{n+1}}$ for $i = 1, 2, \ldots, n$. Then,

$$\sum_{i=1}^{n+1} b_i g(x_i) = b_{n+1} g(x_{n+1}) + (1-b_{n+1}) \sum_{i=1}^{n} q_i g(x_i)$$

$$\geq b_{n+1} g(x_{n+1}) + (1-b_{n+1}) g\left(\sum_{i=1}^{n} q_i x_i\right)$$

from induction hypothesis

$$\geq g\left(b_{n+1} x_{n+1} + (1-b_{n+1}) \sum_{i=1}^{n} q_i x_i\right)$$

from convexity

Now we proceed to prove Proposition 1, i.e., $K(p, p^0) \geq 0$.

$$K(p, p^0) = \sum_{x \in \Theta} p(x) \log \frac{p(x)}{p^0(x)} \quad \text{where } \Theta = \{x \in \Omega : p(x) > 0\}$$

and $\Omega$ is the sample space

(We exclude those points $x \in \Omega$ for which $p(x) = 0$):

$$= \sum_{x \in \Theta} p(x) \left(-\log \frac{p^0(x)}{p(x)}\right) \geq -\log \left(\sum_{x \in \Theta} p(x) \frac{p^0(x)}{p(x)}\right)$$

by Jensen's inequality and noting that $-\log$ is a strictly convex fn.
Therefore,
\[
K(p, p^0) \geq -\log \sum_{x \in \Theta} p^0(x) \\
\geq -\log \sum_{x \in \Omega} p^0(x) \text{ because } \Theta \subseteq \Omega, \\
= -\log(1) \\
= 0
\]

Thus, the inequality \( K(p, p^0) \geq 0 \) is established.

The equality holds if and only if \( \frac{p^0(x)}{p(x)} \text{ is constant } \forall x \in \Theta \) in
Jensen's inequality. However, since \( \sum_{x \in \Theta} p(x) = \sum_{x \in \Theta} p^0(x) = 1 \),
it follows that the constant is 1, i.e. \( p^0(x) = p(x) \forall x \in \Theta \)
For \( x \in \Omega \setminus \Theta \), if \( p^0(x) > 0 \), then also \( K(p, p^0) = 0 \).

**Definition:** Let \( X \) and \( Y \) be two random variables.

Let \( p_{XY}(\theta, \phi) \) be the joint probability mass fn and the
respective marginal probability mass functions be \( p_X(\theta) \)
and \( p_Y(\phi) \). Then, the mutual information between \( X \) and \( Y \)
is defined as:
\[
I(X, Y) = \sum_{\theta} \sum_{\phi} p(\theta, \phi) \log \frac{p(\theta, \phi)}{p_X(\theta) p_Y(\phi)}
\]

**Corollary:** (Non-negativity of mutual information)
\[
I(X, Y) \geq 0 \text{ and equality holds iff } X \text{ and } Y \text{ are independent}
\]

**Proof:** \( I(X, Y) = K(p_{XY}, p_X p_Y) \geq 0 \) and equality
holds if and only if \( p_{XY} = p_X p_Y \Leftrightarrow \) the random
variables \( X \) and \( Y \) are independent.
Proposition 2: If \( p^0 \) is uniform distribution, i.e., \( p^0_i = \frac{1}{l} \) \( \forall i \), then

\[
K(p, p^0) = I(p) + \log(l)
\]

Proof: Given \( p^0_i = \frac{1}{l} \) \( \forall i \),

\[
K(p, p^0) = \sum_i p_i \log(l p_i) = \sum_i p_i \log(p_i) + \log(l) \sum_i p_i = I(p) + \log(l)
\]

Proposition 3: \( K(p, p^0) \) is a convex function if \( p^0 \) is uniformly distributed.

Proof:

\[
\frac{\partial^2 K}{\partial p_i \partial p_j} = \frac{\partial^2}{\partial p_i \partial p_j} \left( \sum_k p_k \log(p_k) + \log(l) \right) = \frac{\partial}{\partial p_i} \left( \log p_i + 1 \right) = \frac{1}{p_i} \delta_{ij} > 0 \Rightarrow \text{convexity}
\]

Proposition 4: Let \( \tilde{\beta} \geq \beta > 0 \). Then,

\[
\frac{\beta^{-1}}{\tilde{\beta}} I_{\tilde{\beta}}(p) \geq \frac{\beta^{-1}}{\beta} I_{\beta}(p) \quad \forall \beta
\]

Proof: We note that \( x^\alpha \) is a convex function of \( x \) for \( \alpha \geq 1 \),

\[
\text{i.e., } (c x + (1-c) \tilde{x})^\alpha \leq c x^\alpha + (1-c) \tilde{x}^\alpha \quad \forall x, \tilde{x}, c \in [0,1]
\]

\( x^\alpha \) is a concave function of \( x \) for \( \alpha \leq 1 \)

\[
\text{i.e., } (c x + (1-c) \tilde{x})^\alpha \geq c x^\alpha + (1-c) \tilde{x}^\alpha \quad \forall x, \tilde{x}, c \in [0,1]
\]

Hence, for arbitrary \( x_i \in [0, \infty) \), \( (\sum_i x_i)^\alpha \begin{cases} \geq \sum_i x_i^\alpha & \text{for } \alpha \geq 1 \\ \leq \sum_i x_i^\alpha & \text{for } \alpha \leq 1 \end{cases} \)

Let \( x_i = (p_i) \beta \) and \( \alpha = \frac{\beta}{\tilde{\beta}} > 1 \). Then,

\[
\left( \sum_i (p_i) \beta \right)^{\frac{\tilde{\beta}}{\beta}} \geq \sum_i (p_i) \tilde{\beta} \Rightarrow \left( \sum_i (p_i) \beta \right)^{\frac{1}{\beta}} \geq \left( \sum_i (p_i) \tilde{\beta} \right)^{\frac{1}{\tilde{\beta}}}
\]

Since \( \log \) is a monotonically increasing function,

\[
\frac{1}{\beta} \ln \left( \sum_i (p_i) \beta \right) \geq \frac{1}{\tilde{\beta}} \ln \left( \sum_i (p_i) \tilde{\beta} \right) \Rightarrow \frac{\beta^{-1}}{\tilde{\beta}} I_{\beta}(p) \geq \frac{\beta^{-1}}{\beta} I_{\beta}(p) \quad \forall \beta \quad \forall \tilde{\beta} \geq \beta > 0
\]

Remark: Proposition 4 provides a lower bound on \( I_{\beta} \) in terms of \( I_{\tilde{\beta}} \).
Proposition 5: Rényi information $I_{\beta}(P)$ is a monotonically increasing function of $\beta$ for any given distribution $P$. That is, $I_{\beta}(P) \geq I_{\beta'}(P)$ if $\beta \geq \beta'$. 

Proof: Differentiating the following expression w.r.t $\beta$

$$I_{\beta}(P) = \frac{1}{\beta - 1} \ln \left( \sum_i P_i^\beta \right) \quad \text{where } P = [p_1, \ldots, p_r]$$

Yields

$$\frac{\partial I_{\beta}}{\partial \beta} = \frac{1}{(\beta - 1)^2} \ln \left( \sum_i P_i^\beta \right) + \frac{\sum (P_i^\beta \ln P_i)}{\sum P_i^\beta}$$

$$= \frac{1}{(\beta - 1)^2} \left[ - \ln \left( \sum_i P_i^\beta \right) + (\beta - 1) \sum P_i^\beta \ln P_i \right]$$

where $P_i = \frac{(P_i)^\beta}{\sum_j (P_j)^\beta}$

$$= \frac{1}{(\beta - 1)^2} \sum P_i \ln \left( \frac{P_i^\beta / \sum P_i^\beta}{P_i} \right)$$

because $\sum P_i = 1$

$$= \frac{1}{(\beta - 1)^2} \sum P_i \ln \frac{P_i}{p_i}$$

Note that $\sum P_i \ln \frac{P_i}{p_i}$ is the Kullback information gain $K(P, P') \geq 0$. Therefore, $\frac{\partial I_{\beta}(P)}{\partial \beta} \geq 0 \forall \beta \forall P$.

Remark: The terms $P_i = \frac{P_i^\beta}{\sum P_i^\beta}$ are called escort probabilities. $P_i \rightarrow p_i \forall i$ as $\beta \rightarrow 1$. The escort probabilities are extensively used in statistical mechanics. The term $\sum P_i^\beta$ is known as the partition function $Z(\beta)$. The term $\beta$ is interpreted as the inverse temperature based on the notation $\beta = \frac{1}{k_B T}$ where $k_B$ is the Boltzmann constant.
Proposition 6:

Define \( \Psi(\beta) \equiv -\log(\sum P_i^{\beta}) = (1-\beta)I_\beta(P) \). Then, \( \Psi(\beta) \) is a monotonically increasing function of \( \beta \) and is concave in \( \beta \), i.e.

\[
\Psi(\tilde{\beta}) \geq \Psi(\beta) \quad \text{for } \tilde{\beta} \geq \beta
\]

and \( \frac{\partial^2 \Psi}{\partial \beta^2} \leq 0 \).

Proof:

\[
\frac{\partial \Psi}{\partial \beta} = -\frac{\sum(b_i^\beta \log b_i)}{\sum b_i^\beta} = -\sum P_i \frac{\log b_i}{b_i} \geq 0 \quad \text{because } b_i \in [0, 1]
\]

\[
\frac{\partial^2 \Psi}{\partial \beta^2} = \frac{(\sum b_i^\beta \log b_i)^2}{\sum b_i^\beta} - \frac{\sum b_i^\beta (\log b_i)^2}{\sum b_i^\beta}
\]

\[
= \left( \sum P_i \log b_i \right)^2 - \sum P_i (\log b_i)^2 < 0 \quad \text{because } 0 \leq P_i \leq 1
\]

Summary of Inequalities Related to Renyi Information

- \( \frac{\partial}{\partial \beta} \left( \frac{\beta-1}{\beta} I_\beta(P) \right) \leq 0 \) (See Proposition 4)

- \( \frac{\partial}{\partial \beta} \left( I_\beta(P) \right) \geq 0 \) (See Proposition 5)

- \( \frac{\partial}{\partial \beta} \left( (1-\beta)I_\beta(P) \right) \geq 0 \) (See Proposition 6)

- \( \frac{\partial^2}{\partial \beta^2} \left( (1-\beta)I_\beta(P) \right) \leq 0 \) (See Proposition 6)
Information-Theoretic Formulation of Canonical Distributions

In statistical mechanics, thermodynamic equilibrium is expressed by using different types of probability distribution:

- Microcanonical
- Canonical ensemble
- Pressure ensemble
- Grand canonical ensemble

Together these probability distributions are referred to as generalized canonical distributions.

Let \( i = 1, 2, \ldots, R \) be a set of microstates. Let \( \tilde{E} \) be a random variable, denoting these microstates, such that the expected value \( \langle \tilde{E} \rangle \equiv \sum_{i=1}^{R} P_i E_i \) where \( P_i \geq 0 \) and \( \sum_{i=1}^{R} P_i = 1 \) (1)

where the realization of the \( i \)th microstate be \( E_i \), whose probability is \( P_i \).

Let us denote \( \langle \tilde{E} \rangle \) as \( E \). Note that the mean value \( E \) could be energy, but not necessarily so.

Let us assume that the thermodynamic system be represented by several random quantities that can be macroscopically observed as \( E^0 \):

\[ E^0 = \sum_{i=1}^{R} P_i E^0_i \] (2)

where \( E^0_i \) is the realization of the 0-component in the \( i \)th state.

By the entropy maximization principle (equivalently, information minimization), the probability vector \( [P_1, \ldots, P_R] \) under thermodynamic equilibrium is attained such that the information \( \sum_i P_i \ln P_i \) (Note: the escort probabilities \( P_i \) become \( P_i \) for \( \beta = 1 \)) is minimized. To see this, let us apply variational principles, where infinitesimal variations \( EP_i \) satisfy Eqs. (1) and (2) as:

\[ \sum_{i=1}^{R} E_i S P_i = 0 \quad \text{and} \quad \sum_{i=1}^{R} S P_i = 0 \] (3)
\[ S_I(P) = \sum_{i=1}^{R} \left[ (P_i + SP_i) \ln(P_i + SP_i) - P_i \ln P_i \right] = 0 \quad (4) \]

For \( |SP_i| \to 0 \quad \forall i \), by Taylor series expansion, Eq. (4) reduces to
\[ S_I(P) = \sum_{i=1}^{R} \left[ \left( 1 + \ln P_i \right) SP_i + o(SP_i) \right] \]
\[ = \sum_{i=1}^{R} \left( 1 + \ln P_i \right) SP_i \quad (5) \]

By using the constraint \( \sum_i SP_i = 0 \) in Eq. (3) to Eq. (5)
\[ S_I(P) = \sum_{i=1}^{R} \left( \ln P_i \right) SP_i \quad (6) \]

Let us use the Lagrange multipliers of \( \beta_0 \) and \( \Psi \) in the equations in (3), i.e., \( \sum_i \beta_0 E_i^\sigma SP_i = 0 \) and \( \sum_i \Psi SP_i = 0 \), in (6) to yield
\[ \sum_{i=1}^{R} \left( \ln P_i - \Psi + \beta_0 E_i^\sigma \right) SP_i = 0 \quad (7) \]

where the tensor notation of repeated symbol’s has been used for the Greek symbol \( \Psi \) to denote summation over all components. Since the equality in (7) is valid for arbitrary \( SP_i \), we have
\[ \ln P_i - \Psi + \beta_0 E_i^\sigma = 0 \]
\[ \Rightarrow P_i = e^{\Psi - \beta_0 E_i^\sigma} \quad (8) \]

Note the Lagrange multiplier coefficients \( \Psi \) and \( \beta_0 \) must be chosen appropriately to fulfill the following requirements:
\[ \sum_{i=1}^{R} P_i = 1 \quad \text{and the macroscopic observation } E^\sigma = \sum_i P_i E_i^\sigma \]

Often the distribution \( P = [P_1 \ldots P_R] \) is called the generalized canonical distribution or Gibbs distribution. \( \Psi \) is called the generalized free energy, \( \beta_0 \)'s are called "intensities" and \( E^\sigma \)'s are called "extensities."
If the mean values $E_i^0$ are not specified, the unbiased guess is the uniform distribution, where $P_i = \frac{1}{R} \forall i \in \{1, \ldots, R\}$. This is what Boltzmann hypothesized, "all microstates are equally probable in the absence of any constraint."

If we define the information $I(P) = \Psi - \beta_0 E_0^0$, and the entropy $S(P) = -I(P)$, then

$$S = -\Psi + \beta_0 E_0^0$$  \hspace{1cm} (9)

Under the normalization constraint

$$\sum_{i=1}^{R} P_i = 1 \Rightarrow \sum_{i=1}^{R} \exp(\Psi - \beta_0 E_0^0) = 1$$  \hspace{1cm} (10)

Then, we must choose

$$\Psi = -\ln Z$$  \hspace{1cm} (11)

where the partition function $Z$ is defined as:

$$Z \equiv \sum_{i=1}^{R} \exp(-\beta_0 E_0^0)$$  \hspace{1cm} (12)

For example, if a system consists of several interacting components, e.g., $U = U(S, V, N_1, \ldots, N_2)$, then we may say that the extensive parameters are denoted as macroscopically observed variables $E_0^0$. The intensive parameters, such as pressure, temperature, and chemical potentials are variables $\beta_0^0$. So, $\beta_0$ represents relative much fewer variables than the number $R$ of microstates.
Thermodynamic Equilibrium Ensembles

Depending on the set of macroscopic thermodynamic observables, denoted as $E^o$, the unbiased guess of probability distribution is interpreted as the following ensembles:

1) **Microcanonical Distribution**: If all $E^o$ are introduced as exact parameters, we assume that each microstate is equally likely. No macroscopic observable is available.

2) **Canonical Distribution**: Only one macroscopic observable, namely some form of energy, is available, i.e. $0 \leq \mathcal{E}_i \leq E^1$, and the system is in thermodynamic equilibrium. Then,

$$P_i = \exp \left( \frac{\mathcal{E}_i - \beta E_i}{k_B T} \right)$$

(13)

Where $\beta = \frac{1}{k_B T}$ where $T$ is the absolute temperature (°K), and Boltzmann constant $k_B = 1.38065 \times 10^{-23}$ Joule/°K.

The generalized free energy $\Psi$ is chosen as: $\Psi(\beta) = \beta F$ where $F$ is the Helmholtz free energy, i.e., $F = \frac{1}{\beta} \Psi(\beta)$.

The resulting partition function is

$$Z(\beta) = \exp (-\beta F) = \sum_i \exp (-\beta E_i)$$

(14)

because $\sum P_i = 1 \Rightarrow \exp (\mathcal{E}_i) \left( \sum_i \exp (-\beta E_i) \right) = 1 \Rightarrow \exp (-\beta \mathcal{E}_i) = \sum_i \exp (-\beta E_i)$

Hence, (13) can be rewritten as

$$P_i = \exp \left[ \frac{\beta(F - E_i)}{k_B T} \right] = \exp \left[ \frac{E - F}{k_B T} \right]$$

(15)

Recall that $F = E - TS$; or $S = \frac{E - F}{T}$. In this case,

$$S = -\ln (\mathcal{P}) = \frac{E - F}{T}$$

(16)
3) **Pressure Ensemble Distribution**: Two macroscopic observables, namely mean values of energy and volume, are available. In the thermodynamic sense, there is a chamber with a movable, diathermal boundary. The volume and energy are fluctuating variables in the sense that they can assume different values in different microstates. Then,

$$P_i = \exp \left[ \frac{\beta (G - E_i - \Pi V_i)}{T} \right]$$  \hspace{1cm} (17)

where $G$ is the Gibbs free energy; $\Pi$ is identified with negative thermodynamic pressure \( \left( \frac{\partial E}{\partial V} \bigg|_{S,N} \right) \); $\beta = \frac{1}{T} = \left( \frac{\partial S}{\partial E} \bigg|_{V,N} \right)$. Now, the probabilities $P_i$ can be expressed as:

$$P_i = \exp \left[ \beta (\Pi V_i - \beta E_i + \Pi V_i) \right] \quad \text{with} \quad \Pi = \beta G$$

Recall that $G = E - TS + \Pi V$ and hence

$$S = \frac{E - \Pi V - G}{T} = \beta \left( E - \Pi V - G \right)$$  \hspace{1cm} (18)

4) **Grand Canonical Ensemble Distribution**: Several macroscopic observables, namely mean values of energy, volume, and mole numbers, are available. In the thermodynamic sense, there is a chamber with a movable, porous, and diathermal boundary. The energy, volume, and mole numbers are fluctuating variables in the sense that they can assume different values in different microstates. Then,

$$P_i = \exp \left[ \frac{\beta (\Omega - E_i - \Pi V_i - \mu N_i)}{T} \right]$$  \hspace{1cm} (19)

where $\mu$ is called chemical potential and $\Omega$ has no special name.

Recall that $\Omega = E - TS - \Pi V - \mu N$ and hence

$$S = \frac{E - \Pi V + \mu N - \Omega}{T}$$  \hspace{1cm} (20)

For a multi-component system $\mu N$ is replaced by $\mu_\alpha N_\alpha$, with $\alpha$ summing over the index $\alpha$. 
Remark: For macroscopic thermodynamic systems, fluctuations of the random quantities are extremely small compared to the respective mean values $E^0$ at the equilibrium condition. Therefore, for equilibrium thermodynamics, $E^0$ can be approximated by $E^0$ in the calculation of the entropy $S$. In the thermodynamic limit, the volume and mole numbers become extremely large, whereas the intensities are kept constant. In this limit, the relations between the macroscopic thermodynamic parameters coincide for different energy representations, provided that entropy $S$ (i.e., the information $I = -S$) has the same value in different representations. In that case, various free energies are related as:

$$G = F - \Pi V = -\Omega + MN$$  \hspace{1cm} (21)

**Principle of Minimum Free Energy Revisited**

The postulate II of the maximum entropy principle starts with given extensities $E^0$. The intensities are determined by the thermodynamic boundary conditions of the system in contact with the environment. By use of Legendre transformation, the fundamental equation in energy representation (or entropy representation) is expressed in terms of the intensities, such as $\beta_0$'s.

Let the probabilities of microstates be denoted as $p_i$ such that $p_i > 0 \forall i$ and $\sum_i p_i = 1$. Let $\delta p_i$ be variations in $p_i$ such that $\sum_i \delta p_i = 0$. The principle of maximum entropy mandates:

$$\delta S = 0 \quad \text{and} \quad \delta E^0 = \sum_i \delta p_i E^0_i = 0$$  \hspace{1cm} (22)

Hence, for fixed values of intensities $\beta_0$, we have:

$$S(S - \beta_0 E^0) = 0$$  \hspace{1cm} (23)
For any probability distribution $\rho$, the generalized free energy

$$\Psi(\rho) = \beta_0 E^\sigma - S = \sum_{i=1}^{\mathcal{R}} \beta_i \beta_0 E^\sigma_i + \sum_{i=1}^{\mathcal{R}} \beta_i \ln \rho_i = \sum_{i=1}^{\mathcal{R}} \beta_i \left( \beta_0 E^\sigma_i + \ln \rho_i \right)$$  \hspace{1cm} (24)

Remark: Generalized Canonical distribution is defined for thermodynamic equilibrium. However, (24) holds also for non-equilibrium conditions and non-thermodynamic systems.

We now proceed to formulate the principle of minimum energy using the variational approach, in which $\Psi(\rho)$ is required to have a minimum in the space of probability distributions.

This minimum is adopted for the canonical distribution:

$$P_i = \exp \left( \Psi - \beta_0 E^\sigma_i \right)$$  \hspace{1cm} (25)

where $\Psi = -\ln Z = -\ln \left( \prod_{i=1}^{\mathcal{R}} \exp \left( -\beta_0 E^\sigma_i \right) \right)$ \hspace{1cm} (26)

Depending on the intensities $\beta_0$, $\Psi$ becomes a specific type of energy. For example,

* If $\beta_0 \sim \frac{1}{T}$ is given and $U$ is not a fluctuating parameter, then
  
  $$\Psi = \beta F$$, i.e. Helmholtz energy $F = \frac{1}{\beta} \Psi$

* If $\beta \sim \frac{1}{T}$ and pressure are given, then
  
  $$\Psi = \beta G$$, i.e. Gibbs energy $G = \frac{1}{\beta} \Psi$

Remark: One can start with Kullback information gain instead of relying on Shannon entropy. In that case, $SS = 0$ in (22) should be replaced by

$$SK(\rho, \rho^0) = \sum_{i=1}^{\mathcal{R}} \left( 1 + \ln \left( \frac{\rho_i}{\rho_i^0} \right) \right) \delta \rho_i = 0$$  \hspace{1cm} (27)

To obtain

$$\rho_i = \rho_i^0 \exp \left( \Psi - \beta_0 E^\sigma_i \right)$$  \hspace{1cm} (28)
Let \( \rho^0 = \exp(\Psi^0 - \beta^0 \mathbf{E}_i^0) \)  

Then,  
\[
K(p, p^0) = \sum_{i=1}^{R} \phi_i \ln \left( \frac{p_i}{p_i^0} \right) - \sum_{i=1}^{R} \left[ p_i \ln p_i - p_i^0 \ln p_i^0 - (p_i - p_i^0) \ln p_i^0 \right]
\]

\[
= \sum_{i=1}^{R} \left[ p_i \ln p_i - p_i^0 \ln p_i^0 - (p_i - p_i^0) \ln p_i^0 \right]
\]

\[
= \sum_{i=1}^{R} \left[ p_i \ln p_i - p_i^0 \ln p_i^0 - (p_i - p_i^0) (\Phi^0 - \beta^0 \mathbf{E}_i^0) \right]
\]

\[
= (S - S^0) + \Psi^0 (\sum \mathbf{E}_i^0 - \sum \mathbf{E}_i^0) + \beta^0 (\mathbf{E}^0 - \mathbf{E}^0)
\]

\[
= -\Delta S + \beta^0 \Delta E^0.
\]

Given the reference (e.g., source) temperature \( T^0 \) and (negative) pressure \( p^0 \),

\[
A \equiv \Delta E - T^0 \Delta S - p^0 \Delta V
\]

is called availability or exergy. In general, if source has defined intensities \( \beta^0 \), then

\[
A = T^0 K(p, p^0) \geq 0 \text{ under equilibrium.}
\]

Because \( K(p, p^0) \geq 0 \) and \( T^0 > 0 \) under equilibrium.

Gibbs Fundamental Equation.

Let us start with the generalized canonical distribution \( p_i \), i.e.,

\[
p_i = \exp(\Psi - \beta_i \mathbf{E}_i) \quad i = 1, 2, \ldots, R
\]

and compute the Kullback information for a perturbed distribution \( p_i + \delta p_i \), which is a consequence of a variation \( \delta \beta_i \) of the intensities \( \beta_i \).

\[
0 \leq K(p + \delta p, p) = \sum_{i} \left( p_i + \delta p_i \right) \ln \left( \frac{p_i + \delta p_i}{p_i} \right)
\]

\[
= \sum_{i} \left[ (p_i + \delta p_i) \ln (p_i + \delta p_i) - p_i \ln (p_i) - \delta p_i \ln (p_i) \right]
\]

Because \( \sum \delta p_i = 0 \)

\[
\delta E_i = \sum \mathbf{E}_i \delta p_i
\]

and \( (\sum \mathbf{E}_i) \delta p_i = 0 \)

\[
\delta E_i = \sum \mathbf{E}_i \delta p_i
\]

\[
\Rightarrow - \delta S + \beta \delta E \geq 0
\]

(32)
If the variations are infinitely small, then
\[ \delta S = \frac{\partial S}{\partial E^\sigma} \delta E^\sigma + \frac{1}{2} \frac{\partial^2 S}{\partial E^\sigma \partial E^\tau} \delta E^\sigma \delta E^\tau \] (33)

Using (33) in (32) yields
\[ \left( \frac{\partial S}{\partial E^\sigma} - \beta_0 \right) \delta E^\sigma + \frac{1}{2} \frac{\partial^2 S}{\partial E^\sigma \partial E^\tau} \delta E^\sigma \delta E^\tau \leq 0 \] (34)

Since the magnitude and sign of \( \delta E^\sigma \) and \( \delta E^\tau \) are arbitrary, we must have the first term on the left-hand side of (34) must be zero, i.e.,
\[ \frac{\partial S}{\partial E^\sigma} = \beta_0 \] (35)

and negativity of the second term and usage of (35) yield
\[ 0 \geq \frac{\partial^2 S}{\partial E^\sigma \partial E^\tau} \delta E^\sigma \delta E^\tau = \frac{\partial \beta_0}{\partial E^\tau} \delta E^\sigma \delta E^\tau = \delta E^\sigma \delta \beta_0 \] (36)

In view of (35), we claim that
\[ S(E) = -\langle \ln \mathcal{P} \rangle = -\Psi + \beta_0 E^\sigma \]

is the Legendre transform of \( \Psi(\beta) \) which yields
\[ \frac{\partial \Psi}{\partial \beta_0} = E^\sigma \] (holding \( S \) constant) (37)

Recalling that, for grand canonical ensemble with fluctuating volume, the general equation is
\[ dS = \beta_0 dE^\sigma \Rightarrow dE = Tds + \Pi dV + \mu dN \] (38)

This is known as the Gibbs fundamental equation, where extensivities are \( S, V \) and \( N \), and intensities are \( T, \Pi \) and \( \mu \).
Susceptibilities and Fluctuations

Let us examine the second order inequality in (36) in terms of the variations $S\beta$

$$0 \geq SE^0 S\beta^0 = \frac{2E^0}{\beta E} S\beta^2 S\beta^0$$

Define susceptibility matrix as:

$$Q = \left[ \frac{2E^0}{\beta E} \right]$$

(41)

where $Q^{\alpha \gamma} = \frac{2E^0}{\beta E}$ is called a susceptibility element.

Imagine the concept of susceptibility with that in electromagnetics.

In free space, permeability $\mu_0$ (Henry/meter) is obtained as

$$\mu_0 = \frac{\beta B}{\beta H} \Rightarrow \beta B = \frac{\beta H}{\mu_0} \beta B$$

where $B$ is flux density (weber/m²) and $H$ is magnetic field intensity (ampere/m). In ferromagnetic media

$$S H = \frac{2}{\mu_0} S B - S M \Rightarrow S B = \mu_0 (1 + \frac{\beta M}{\beta H}) S H$$

where $M$ is magnetization (i.e., magnetic dipole moment per unit volume). The dimensionless quantity

$$\chi = \frac{\beta M}{\beta H}$$

is called magnetic susceptibility and $\mu = \mu_0 (1 + \chi)$.

Similarly, in electrostatics, permittivity $\varepsilon$, is defined as

$$\varepsilon = \varepsilon_0 (1 + \chi) \text{ in units of farads/meter}$$

where $\varepsilon$ is permittivity of a dielectric material with

$$D = \varepsilon E$$

where $D =$ electric flux density (Coulomb/m²)

$E =$ electric field intensity (volts/m)

$$\chi = \frac{1}{\varepsilon_0} \frac{\partial P}{\partial E} \text{ where } P = \text{polarization or electric dipole moment/} \text{unit volume (Coulomb/m²)}$$
It also follows from (36) that
\[ 0 \leq -8\beta_0\Delta E^0 = -\frac{\partial \beta_0}{\partial E^0} \Delta E^0 \Delta E^0 \]  
(42)

We also refer to the following term as the susceptibility matrix
\[ R = \left[ -\frac{\partial \beta_0}{\partial E^0} \right] \]  
(43)

where \( R_{\alpha \alpha} = -\frac{\partial \beta_0}{\partial E^0} \) is called a susceptibility element.

Note that the elements of \( Q \) and \( R \) are related as:
\[ Q_{\alpha \beta}^0 \equiv R_{\alpha \beta} = \frac{\partial E^0}{\partial \beta_0} \frac{\partial \beta_\alpha}{\partial E^0} = \frac{\partial E^\alpha}{\partial E^0} = \delta_{\alpha \beta} \]  
(44)

Next, let us bring in the notion of fluctuations. Differentiation of the expression with respect to \( \beta_\alpha \)
\[ E^\alpha = \sum_i P_i \ v_i E_i^\alpha = \sum_i \exp(\varepsilon_i - \beta_\alpha E_i^\alpha) E_i^\alpha \]  
(45)

yields the susceptibilities
\[ Q_{\alpha \beta}^0 = \sum_i \left[ \frac{\partial E_i^\alpha}{\partial \beta_\alpha} - E_i^\alpha \right] P_i E_i^\beta = \sum_i \left( E_i^\alpha - E_i^\beta \right) P_i E_i^\alpha \]

\[ \Rightarrow Q_{\alpha \beta}^0 = \langle \tilde{E}_i^\alpha \tilde{E}_i^\beta \rangle - \langle \tilde{E}_i^\alpha \rangle \langle \tilde{E}_i^\beta \rangle \]  
(46)

Defining the fluctuations of \( E_i^\alpha \) as
\[ \Delta E_i^\alpha = E_i^\alpha - E_i^\alpha = \langle \tilde{E}_i^\alpha \rangle - E_i^\alpha \]  
(47)

we obtain
\[ \langle \Delta E_i^\alpha \Delta E_i^\beta \rangle = \langle \tilde{E}_i^\alpha \tilde{E}_i^\beta \rangle - \langle \tilde{E}_i^\alpha \rangle \langle \tilde{E}_i^\beta \rangle = Q_{\alpha \beta}^0 \]  
(48)

Hence, the susceptibility matrix is also the correlation matrix.

Now, we can express (40) and (42) as:
(Concave function of extensities)
\[ S^2 S = \frac{\partial^2 S}{\partial E^0 \partial E^0} S E^0 S E^0 \leq 0 \]  
(49)

(Concave function of intensities)
\[ S^2 \Phi = \frac{\partial^2 \Phi}{\partial \beta^0 \partial \beta^0} S \beta^0 S \beta^0 \leq 0 \]  
(50)