CHAPTER I

ENTROPY

In this chapter, basic ideas of entropy are presented from the work of Shannon and Kolmogorov-Sinai. The first one is defined for finite schema and the second is for measure preserving transformations. Conjugacy between two measure preserving transformations is considered in terms of their algebraic models. When a transformation is fixed, the entropy is defined for all transformation invariant probability measures. In this case, it is called an entropy functional. An integral representation of this functional is given. Relative entropy and Kullback-Leibler information are studied in connection with sufficient statistics and hypothesis testing.

1.1. The Shannon entropy

We consider basic properties and axioms of Shannon's entropy. Let $n \in \mathbb{N}$ (the set of all positive integers) and $X = \{x_1, \dots, x_n\}$ be a finite set with a probability distribution $\mathbf{p} = (p_1, \dots, p_n)$, i.e., $p_j = p(x_j) \geq 0$, $1 \leq j \leq n$ and $\sum\limits_{j=1}^{n} p_j = 1$, where $p(\cdot)$ denotes the probability. We usually denote this as $(X, \mathbf{p})$ and call it a **complete system of events** or a **finite scheme**. The **entropy** or the **Shannon entropy** $H(X)$ of a finite scheme $(X, \mathbf{p})$ is defined by

$$H(X) = -\sum_{j=1}^{n} p_j \log p_j, \tag{1.1}$$

where "log" is the natural logarithm and we regard $0 \log 0 = 0 \log \frac{0}{0} = 0$. We also say that $H(X)$ is the **uncertainty** or **information** of the system $(X, \mathbf{p})$. Justification of these terminologies will be clarified later in this section. Since RHS (= right hand side) of (1.1) depends only on the probability distribution $\mathbf{p} = (p_1, \dots, p_n)$ we may also write

$$H(X) = H(\mathbf{p}) = H(p_1, \dots, p_n) = -\sum_{j=1}^{n} p_j \log p_j.$$

We need some notations. For $n \in \mathbb{N}$, let Δ_n denote the set of all n-dimensional probability distributions $\mathbf{p} = (p_1, \dots, p_n)$, i.e.,

$$\Delta_n = \left\{ \mathbf{p} = (p_1, \dots, p_n) : \sum_{j=1}^{n} p_j = 1,\, p_j \geq 0,\, 1 \leq j \leq n \right\}.$$

Let $Y = \{y_1, \dots, y_m\}$ be another finite set. The probability of (x_j, y_k) and the conditional probability of x_j given y_k are respectively denoted by $p(x_j, y_k)$ and $p(x_j|y_k) = \frac{p(x_j, y_k)}{p(y_k)}$ if $p(y_k) > 0$. Then the **conditional entropy** $H(X|Y)$ of X given Y is defined by

$$H(X|Y) = -\sum_{y \in Y} \sum_{x \in X} p(y) p(x|y) \log p(x|y). \tag{1.2}$$

If we define $H(X|y)$, called the **conditional entropy** of X given $Y = y$, by

$$H(X|y) = -\sum_{x \in X} p(x|y) \log p(x|y),$$

then (1.2) is interpreted as the average of these conditional entropies over Y. The quantity $I(X, Y)$ defined below is called the **mutual information** between $(X, \mathbf{p})$ and $(Y, \mathbf{q})$:

$$I(X, Y) = H(X) - H(X|Y)$$

since we can easily verify that

$$\begin{aligned} I(X, Y) &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X, Y) \\ &= \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \geq 0, \end{aligned} \tag{1.3}$$

where

$$H(X, Y) = -\sum_{x,y} p(x, y) \log p(x, y)$$

is the entropy of the **compound scheme** $\big((X, \mathbf{p}), (Y, \mathbf{q})\big)$. The inequality (1.3) will be proved in Theorem 1 below. If we consider two probability distributions $\mathbf{p}, \mathbf{q} \in \Delta_n$ of X, then the **relative entropy** $H(\mathbf{p}|\mathbf{q})$ of $\mathbf{p}$ with respect to $\mathbf{q}$ is given by

$$H(\mathbf{p}|\mathbf{q}) = \sum_{j=1}^{n} p_j (\log p_j - \log q_j) = \sum_{j=1}^{n} p_j \log \frac{p_j}{q_j}.$$

If $p_j > 0$ and $q_j = 0$ for some j, then we define $H(\mathbf{p}|\mathbf{q}) = \infty$. Observe the difference between $H(X|Y)$ and $H(\mathbf{p}|\mathbf{q})$. Relative entropy will be discussed in detail in a later section.

The next two theorems give basic properties of entropies.

Theorem 1. *Consider entropies on* Δ_n.

(1) $H(\mathbf{p}|\mathbf{q}) \geq 0$ *for* $\mathbf{p}, \mathbf{q} \in \Delta_n$, *and* $H(\mathbf{p}|\mathbf{q}) = 0$ *if and only if* $\mathbf{p} = \mathbf{q}$.

(2) *Let* $\mathbf{p} \in \Delta_n$ *and* $A = (a_{jk})$ *be an* $n \times n$ *doubly stochastic matrix, i.e.,* $a_{jk} \geq 0$, $\sum\limits_{j=1}^{n} a_{jk} = \sum\limits_{k=1}^{n} a_{jk} = 1$ *for* $1 \leq j, k \leq n$. *Then,* $\mathbf{q} = A\mathbf{p} \in \Delta_n$ *and* $H(\mathbf{q}) \geq H(\mathbf{p})$. *The equality holds if and only if* $q_k = p_{\pi(k)}$, $1 \leq k \leq n$ *for some permutation* π *of* $\{1, \ldots, n\}$.

(3) $H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$.

(4) $H(X, Y) \leq H(X) + H(Y)$. *The equality holds if and only if* X *and* Y *are independent.*

(5) $H(X|Y) \leq H(X)$. *The equality holds if and only if* X *and* Y *are independent.*

Proof. (1) Assume $H(\mathbf{p}|\mathbf{q}) < \infty$. Using an inequality $t \log t \geq t - 1$ for $t > 0$, we get

$$\frac{p_j}{q_j} \log \frac{p_j}{q_j} \geq \frac{p_j}{q_j} - 1 \qquad \text{or} \qquad p_j \log \frac{p_j}{q_j} \geq p_j - q_j$$

for $j = 1, \ldots, n$. Hence

$$H(\mathbf{p}|\mathbf{q}) = \sum_{j=1}^{n} p_j \log \frac{p_j}{q_j} \geq \sum_{j=1}^{n} (p_j - q_j) = 0.$$

The statement about the equality follows from the fact that $t \log t = t - 1$ if and only if $t = 1$.

(2) $\mathbf{q} = A\mathbf{p} \in \Delta_n$ is clear. Since the function $\phi(t) = -t \log t$ is **concave** (= concave downward) for $t > 0$, we have

$$\begin{aligned} H(\mathbf{q}) &= \sum_{j=1}^{n} \phi(q_j) = \sum_{j=1}^{n} \phi\Big(\sum_{k=1}^{n} a_{jk} p_k\Big) \\ &\geq \sum_{j=1}^{n} \sum_{k=1}^{n} a_{jk} \phi(p_k) = \sum_{k=1}^{n} \phi(p_k) = H(\mathbf{p}). \end{aligned}$$

The equality holds if and only if $\phi\Big(\sum\limits_{k=1}^{n} a_{jk} p_k\Big) = \sum\limits_{k=1}^{n} \phi(p_k)$ for $1 \leq j \leq n$ if and only if for each $j = 1, \ldots, n$, $a_{jk} = 1$ for some k and $a_{jk} = 0$ otherwise if and only if $q_k = p_{\pi(k)}$, $1 \leq k \leq n$ for some permutation π of $\{1, \ldots, n\}$.

(3) Observe the following computations:

$$\begin{aligned} H(X,Y) &= -\sum_{x,y} p(x,y)\log p(x,y) \\ &= -\sum_{x,y} p(x,y)\log p(x)p(y|x) \\ &= -\sum_{x,y} p(x,y)\log p(x) - \sum_{x,y} p(x,y)\log p(y|x) \\ &= H(X) + H(Y|X), \end{aligned}$$

giving the first equality and, similarly, $H(X,Y) = H(Y) + H(X|Y)$.

(4) is derived as follows:

$$\begin{aligned} H(X)+H(Y) &= -\sum_{x} p(x)\log p(x) - \sum_{y} p(y)\log p(y) \\ &= -\sum_{x,y} p(x,y)\log p(x)p(y) \\ &\geq -\sum_{x,y} p(x,y)\log p(x,y), \quad \text{by (1)}, \\ &= H(X,Y). \end{aligned}$$

By (1) the equality holds if and only if $p(x,y) = p(x)p(y)$ for $x \in X$ and $y \in Y$, i.e., X and Y are independent.

(5) is clear from (3) and (4). □

Let $\mathbb{R} = (-\infty, \infty)$, $\mathbb{R}^+ = [0,\infty)$ and $\overline{\mathbb{R}^+} = [0,\infty]$.

Theorem 2. *Let* $\mathbf{p} = (p_j)$, $\mathbf{q} = (q_j) \in \bigcup\limits_{n=2}^{\infty} \Delta_n$.

(1) (**Positivity**) $H(\mathbf{p}) \geq 0$.

(2) (**Continuity**) $H : \bigcup\limits_{n=2}^{\infty} \Delta_n \to \overline{\mathbb{R}^+}$ *is continuous.*

(3) (**Monotonicity**) $f(n) \equiv H\big(\frac{1}{n}, \dots, \frac{1}{n}\big)$ *is an increasing function of* n *and*

$$H(p_1, \dots, p_n) \leq H\left(\frac{1}{n}, \dots, \frac{1}{n}\right) = f(n).$$

(4) (**Extendability**) $H(p_1, \dots, p_n) = H(p_1, \dots, p_n, 0)$.

(5) (**Symmetry**) $H(p_1, \dots, p_n) = H(p_{\pi(1)}, \dots, p_{\pi(n)})$ *for every permutation* π *of* $\{1, \dots, n\}$.

(6) (**Additivity**)

$$\begin{aligned} &H(p_1q_1, \dots, p_1q_m, p_2q_1, \dots, p_2q_m, \dots, p_nq_1, \dots, p_nq_m) \\ &\qquad\qquad\qquad = H(p_1, \dots, p_n) + H(q_1, \dots, q_m). \end{aligned}$$

(7) (**Subadditivity**) *If* $r_{jk} \ge 0$, $\sum\limits_{j,k} r_{jk} = 1$, $\sum\limits_{k} r_{jk} = p_j$, $\sum\limits_{j} r_{jk} = q_k$, *then*

$$H(r_{11}, \dots, r_{nm}) \le H(p_1, \dots, p_n) + H(q_1, \dots, q_m).$$

(8) (**Concavity**) *For* $\mathbf{p}, \mathbf{q} \in \Delta_n$ *and* $\alpha \in (0,1)$ *it holds that*

$$H\big(\alpha\mathbf{p} + (1-\alpha)\mathbf{q}\big) \ge \alpha H(\mathbf{p}) + (1-\alpha)H(\mathbf{q}).$$

Proof. (1), (2), (4) and (5) are obvious.

(3) $f(n) = \log n$ for $n \ge 1$, so that f is an increasing function. As to the second statement, without loss of generality we can assume $p_j > 0$ for all j. Then

$$\begin{aligned} H(p_1, \dots, p_n) - H\left(\frac{1}{n}, \dots, \frac{1}{n}\right) &= -\log n - \sum_{j=1}^{n} p_j \log p_j \\ &= \sum_{j=1}^{n} p_j \log\left(\frac{1}{np_j}\right) \\ &\le \sum_{j=1}^{n} p_j \left(\frac{1}{np_j} - 1\right), \quad \text{by } t - 1 \ge \log t,\ t > 0, \\ &= 0, \end{aligned}$$

or $H(p_1, \dots, p_n) \le H\big(\frac{1}{n}, \dots, \frac{1}{n}\big)$.

(6) follows from the following computation:

$$\begin{aligned} &H(p_1q_1, \dots, p_1q_m, p_2q_1, \dots, p_2q_m, \dots, p_nq_1, \dots, p_nq_m) \\ &\qquad = -\sum_{j=1}^{n}\sum_{k=1}^{m} p_jq_k \log p_jq_k \\ &\qquad = -\sum_{j,k} p_jq_k \log p_j - \sum_{j,k} p_jq_k \log q_k \\ &\qquad = H(p_1, \dots, p_n) + H(q_1, \dots, q_m). \end{aligned}$$

(7) is a reformulation of Theorem 1 (4).

(8) Since $\phi(t) = -t\log t$ is concave for $t > 0$ we have

$$\phi\big(\alpha p_j + (1-\alpha)q_j\big) \ge \alpha\phi(p_j) + (1-\alpha)\phi(q_j), \qquad 1 \le j \le n.$$

Summing with respect to j we obtain the desired inequality. □

We note that the relative entropy $H(\mathbf{p}|\mathbf{q})$ also has concavity. That is, for $\mathbf{p}_1$, $\mathbf{p}_2$, $\mathbf{q}_1$, $\mathbf{q}_2 \in \Delta_n$ and $\alpha \in (0,1)$ we have that

$$H\big(\alpha\mathbf{p}_1 + (1-\alpha)\mathbf{p}_2\big|\alpha\mathbf{q}_1 + (1-\alpha)\mathbf{q}_2\big) \le \alpha H(\mathbf{p}_1|\mathbf{q}_1) + (1-\alpha)H(\mathbf{p}_2|\mathbf{q}_2). \tag{1.4}$$

A generalized version of this is shown in Theorem 6.3 (2). Here we give an elementary proof to (1.4). First we prove that for a_i, $b_i \ge 0$, $1 \le i \le n$

$$\sum_{i=1}^n a_i \log\frac{a_i}{b_i} \ge \left(\sum_{i=1}^n a_i\right)\log\frac{\sum\limits_{i=1}^n a_i}{\sum\limits_{i=1}^n b_i}, \tag{1.5}$$

where the equality holds if and only if $\frac{a_i}{b_i} = \text{const}$.

In fact, we can assume a_i, $b_i > 0$ for $1 \le i \le n$. Let $\alpha_i = \frac{b_i}{\sum\limits_{j=1}^n b_j}$ and $t_i = \frac{a_i}{b_i}$ for $1 \le i \le n$. Then, since $\alpha_i > 0$ $(1 \le i \le n)$ and $\sum\limits_{i=1}^n \alpha_i = 1$, **Jensen's inequality** for $\phi(t) = t\log t$, $t > 0$

$$\sum_{i=1}^n \alpha_i\phi(t_i) \ge \phi\bigg(\sum_{i=1}^n \alpha_i t_i\bigg)$$

yields that

$$\sum_{i=1}^n \frac{b_i}{\sum\limits_{j=1}^n b_j}\cdot\frac{a_i}{b_i}\log\frac{a_i}{b_i} \ge \sum_{i=1}^n \frac{b_i}{\sum\limits_{j=1}^n b_j}\log\left(\sum_{i=1}^n \frac{b_i}{\sum\limits_{j=1}^n b_j}\cdot\frac{a_i}{b_i}\right).$$

Then the desired inequality (1.5) follows from the above inequality. Now, if write $\mathbf{p}_i = (p_{i1}, \dots, p_{in})$ and $\mathbf{q}_i = (q_{i1}, \dots, q_{in})$ for $i = 1, 2$, then (1.5) implies that for $1 \le j \le n$

$$\big(\alpha p_{1j} + (1-\alpha)p_{2j}\big)\log\frac{\alpha p_{1j} + (1-\alpha)p_{2j}}{\alpha q_{1j} + (1-\alpha)q_{2j}} \le \alpha p_{1j}\log\frac{\alpha p_{1j}}{\alpha q_{1j}} + (1-\alpha)p_{2j}\log\frac{(1-\alpha)p_{2j}}{(1-\alpha)q_{2j}}.$$

Adding both sides with respect to j gives (1.4).

Before characterizing the Shannon entropy we consider the function $f(n) = H\big(\frac{1}{n}, \ldots, \frac{1}{n}\big)$, $n \geq 1$. $f(n)$ stands for the entropy or uncertainty or information that a finite scheme $\big(X, \big(\frac{1}{n}, \ldots, \frac{1}{n}\big)\big)$ has. We impose some conditions on the function $f(n)$. In the case where $n = 1$, there is no uncertainty, so that we have

1°) $f(1) = 0$.

If $n \geq m$, then $\mathbf{p} = \big(\frac{1}{n}, \ldots, \frac{1}{n}\big)$ has more uncertainty than $\mathbf{q} = \big(\frac{1}{m}, \ldots, \frac{1}{m}\big)$. Hence,

2°) $f(n) \geq f(m)$ if $n \geq m$, i.e., f is nondecreasing.

If $\big(X, \big(\frac{1}{n}, \ldots, \frac{1}{n}\big)\big)$ and $\big(Y, \big(\frac{1}{m}, \ldots, \frac{1}{m}\big)\big)$ are two independent schema, the compound scheme is $\big(X \times Y, \big(\frac{1}{nm}, \ldots, \frac{1}{nm}\big)\big)$. In this case, the uncertainty of $X \times Y$ should be equal to the sum of those of X and Y, i.e.,

3°) $f(nm) = f(n) + f(m)$.

Under these conditions we can characterize f as follows.

Proposition 3. *Let $f : \mathbb{N} \to \mathbb{R}^+$ be a function satisfying conditions* 1°), 2°) *and* 3°) *above. Then there exists some $\lambda > 0$ such that*

$$f(n) = \lambda \log n, \qquad n \in \mathbb{N}.$$

Proof. This is well-known in functional equation theory. For the sake of completeness we sketch the proof. By 3°) we have $f(n^2) = 2f(n)$ and, in general,

$$f(n^r) = rf(n), \qquad n, r \in \mathbb{N}, \tag{1.6}$$

which can be verified by mathematical induction. Now let $r, s, n \in \mathbb{N}$ be such that $r, s \geq 2$. Choose $m \in \mathbb{N}$ so that

$$r^m \leq s^n < r^{m+1}.$$

Then

$$m \log r \leq n \log s < (m+1) \log r$$

and hence

$$\frac{m}{n} \leq \frac{\log s}{\log r} < \frac{m}{n} + \frac{1}{n}. \tag{1.7}$$

On the other hand, by 2°) we get

$$f(r^m) \leq f(s^n) < f(r^{m+1})$$

and hence by (1.6)

$$mf(r) \leq nf(s) \leq (m+1)f(r),$$

so that

$$\frac{m}{n} \le \frac{f(s)}{f(r)} \le \frac{m}{n} + \frac{1}{n}. \tag{1.8}$$

Thus (1.7) and (1.8) give

$$\left| \frac{f(s)}{f(r)} - \frac{\log s}{\log r} \right| \le \frac{2}{n}, \qquad n \ge 1,$$

which implies that

$$\frac{f(s)}{\log s} = \frac{f(r)}{\log r}.$$

Since $r, s \ge 2$ are arbitrary, it follows that for some constant $\lambda > 0$

$$f(n) = \lambda \log n, \qquad n \ge 2,$$

and by 1°) the above equality is true for $n = 1$ too. □

Consider a finite scheme $(X, \mathbf{p})$ with $\mathbf{p} = (p_1, \dots, p_n) \in \Delta_n$. If $p(x_j) = p_j = \frac{1}{n}$, then x_j has $\log n = -\log \frac{1}{n}$ as information or entropy, which is justified by Proposition 3. This suggests that each x_j has information of $-\log p_j$ and $H(X) = -\sum\limits_{j=1}^{n} p_j \log p_j$ is the average information that $X = \{x_1, \dots, x_n\}$ has, giving a good reason to define the entropy of $(X, \mathbf{p})$ by (1.1).

To characterize the Shannon entropy we consider two axioms.

The Shannon-Khinchin Axiom.

(1°) $H : \bigcup\limits_{n=2}^{\infty} \Delta_n \to \mathbb{R}^+$ is continuous and, for every $n \ge 2$,

$$H\left(\frac{1}{n}, \dots, \frac{1}{n}\right) = \max \{ H(\mathbf{p}) : \mathbf{p} \in \Delta_n \}.$$

(2°) For every $n \ge 2$ and $(p_1, \dots, p_n) \in \Delta_n$

$$H(p_1, \dots, p_n, 0) = H(p_1, \dots, p_n).$$

(3°) If $\mathbf{p} = (p_1, \dots, p_n) \in \Delta_n$, $p_j = \sum\limits_{k=1}^{m_j} q_{jk}$, $q_{jk} \ge 0$ $1 \le k \le m_j$, $1 \le j \le n$,

$$\begin{aligned} &H(q_{11}, \dots, q_{1m_1}, \dots, q_{n1}, \dots, q_{nm_n}) \\ &\qquad = H(p_1, \dots, p_n) + \sum_{j=1}^{n} p_j H\left(\frac{q_{j1}}{p_j}, \dots, \frac{q_{jm_j}}{p_j} \right). \end{aligned}$$

The Faddeev Axiom.

[1°] $f(p) = H(p, 1-p) : [0,1] \to \mathbb{R}$ is continuous and $f(p_0) > 0$ for some $p_0 \in [0,1]$.

[2°] $H(p_1, \dots, p_n) = H(p_{\pi(1)}, \dots, p_{\pi(n)})$ for every $(p_1, \dots, p_n) \in \Delta_n$ and permutation π of $\{1, \dots, n\}$.

[3°] If $(p_1, \dots, p_n) \in \Delta_n$ and $p_n = q + r > 0$ with q, $r \geq 0$, then

$$H(p_1, \dots, p_{n-1}, q, r) = H(p_1, \dots, p_n) + p_n H\left(\frac{q}{p_n}, \frac{r}{p_n}\right).$$

The Faddeev Axiom is an improvement of the Shannon-Khinchin Axiom since [1°] and [3°] are simpler than (1°) and (3°), and [2°] is very natural. These two axioms are equivalent and they imply the Shannon entropy within a positive constant multiple as is seen in the following theorem.

Theorem 4. *The following statements are equivalent to each other.*

(1) $H(\cdot) : \bigcup\limits_{n=2}^{\infty} \Delta_n \to \mathbb{R}^+$ *satisfies the Shannon-Khinchin Axiom.*

(2) $H(\cdot) : \bigcup\limits_{n=2}^{\infty} \Delta_n \to \mathbb{R}^+$ *satisfies the Faddeev Axiom.*

(3) *There is some* $\lambda > 0$ *such that*

$$H(p_1, \dots, p_n) = -\lambda \sum_{j=1}^{n} p_j \log p_j, \qquad (p_1, \dots, p_n) \in \Delta_n, \ n \geq 2. \tag{1.9}$$

Proof. (1) ⇒ (2). Assume that (1) is true. [1°] follows from (1°).

[2°] is derived as follows. If $p_1, \dots, p_n$ are positive rationals, then $p_j = \frac{\ell_j}{m}$ for some $\ell_1, \dots, \ell_n, m \in \mathbb{N}$. Hence

$$\begin{aligned} H(p_1, \dots, p_n) &= H\left(\frac{\ell_1}{m}, \dots, \frac{\ell_n}{m}\right) \\ &= H\Bigl(\underbrace{\frac{1}{m}, \dots, \frac{1}{m}}_{\ell_1}, \dots, \underbrace{\frac{1}{m}, \dots, \frac{1}{m}}_{\ell_n}\Bigr) - \sum_{j=1}^{n} p_j H\left(\frac{1}{\ell_j}, \dots, \frac{1}{\ell_j}\right). \end{aligned}$$

Thus, for any permutation π of $\{1, \dots, n\}$, $H(p_1, \dots, p_n) = H(p_{\pi(1)}, \dots, p_{\pi(n)})$. The case where p_j's are not necessarily rational follows from the continuity of H ((1°)) and the approximation by sequences of rational numbers.

[3°]. It follows from (2°), (3°) and [2°] that

$$H\left(\frac{1}{2}, \frac{1}{2}\right) = H\left(\frac{1}{2}, \frac{1}{2}, 0, 0\right) = H\left(\frac{1}{2}, 0, \frac{1}{2}, 0\right)$$

$$= H\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{1}{2}H(1,0) + \frac{1}{2}H(1,0),$$

implying $H(1,0) = 0$. Hence

$$\begin{aligned} H(p_1, \ldots, p_n) &= H(p_1, 0, \ldots, p_{n-1}, 0, q, r) \\ &= H(p_1, \ldots, p_n) + \sum_{j=1}^{n-1} p_j H(1,0) + p_n H\left(\frac{q}{p_n}, \frac{r}{p_n}\right) \\ &= H(p_1, \ldots, p_n) + p_n H\left(\frac{q}{p_n}, \frac{r}{p_n}\right), \end{aligned}$$

i.e., [3°] holds.

(2) ⇒ (3). Assume that (2) is true. Using [3°], we have for any $p, q \geq 0$, $r > 0$ with $p + q + r = 1$

$$\begin{aligned} H(p, q, r) &= H(p, q + r) + (q + r)H\left(\frac{q}{q+r}, \frac{r}{q+r}\right) \\ &= H(q, p + r) + (p + r)H\left(\frac{p}{p+r}, \frac{r}{p+r}\right). \end{aligned}$$

If we set $f(p) = H(p, 1-p)$, then the second of the above equalities becomes

$$f(p) + (1-p)f\left(\frac{q}{1-p}\right) = f(q) + (1-q)f\left(\frac{p}{1-q}\right). \tag{1.10}$$

Letting $p = 0$ and $0 < q < 1$, we get

$$f(0) = H(0,1) = 0.$$

Integrating (1.10) with respect to q from 0 to $1-p$ gives

$$(1-p)f(p) + (1-p)^2 \int_0^1 f(t)\,dt = \int_0^{1-p} f(t)\,dt + p^2 \int_p^1 \frac{f(t)}{t^3}\,dt. \tag{1.11}$$

Since $f(p)$ is continuous and hence all terms except the first on the LHS (= left hand side) of (1.11) are differentiable, we see that $f(p)$ is also differentiable on $(0,1)$. By differentiating (1.11) with respect to p we obtain

$$(1-p)f'(p) - f(p) - 2(1-p)\int_0^1 f(t)\,dt = -f(1-p) + 2p\int_p^1 \frac{f(t)}{t^3}\,dt - \frac{f(p)}{p}.$$

We can simplify the above by using $f(p) = f(1-p)$ to get

$$(1-p)f'(p) = 2(1-p)\int_0^1 f(t)\,dt + 2p\int_p^1 \frac{f(t)}{t^3}\,dt - \frac{f(p)}{p}. \tag{1.12}$$

It follows that $f'(p)$ is also differentiable on $(0,1)$. By differentiating (1.12) we have

$$f''(p) = -\frac{2}{p(1-p)}\int_0^1 f(t)\,dt, \qquad 0 < p < 1 \tag{1.13}$$

and integrating (1.13) twice gives

$$f(p) = \alpha p + \beta - 2\bigl\{p\log p + (1-p)\log(1-p)\bigr\}\int_0^1 f(t)\,dt, \tag{1.14}$$

where $\alpha, \beta \in \mathbb{R}$ are constants. Note that $\alpha = 0$ since $f(p) = f(1-p)$ and that (1.14) holds for $0 \le p \le 1$. Thus $\beta = 0$ since $f(0) = 0$. Consequently, letting $\lambda = 2\int_0^1 f(t)\,dt$, it holds that

$$H(p_1, p_2) = -\lambda(p_1\log p_1 + p_2\log p_2), \qquad (p_1, p_2) \in \Delta_2,$$

proving (1.9) for the case $n = 2$.

For a general $n \ge 2$ we prove (1.9) by mathematical induction. Suppose that

$$H(p_1, \dots, p_n) = -\lambda\sum_{j=1}^n p_j\log p_j, \qquad (p_1, \dots, p_n) \in \Delta_n$$

and take any $(q_1, \dots, q_{n+1}) \in \Delta_{n+1}$. We can assume that $q_{n+1} > 0$. Then observe that

$$\begin{aligned} H(q_1, \dots, q_n, q_{n+1}) &= H(q_1, \dots, q_{n-1}, q_n + q_{n+1}) \\ &\qquad + (q_n + q_{n+1})H\left(\frac{q_n}{q_n + q_{n+1}}, \frac{q_{n+1}}{q_n + q_{n+1}}\right) \\ &= -\lambda\sum_{j=1}^{n-1} q_j\log q_j - \lambda(q_n + q_{n+1})\log(q_n + q_{n+1}) \\ &\qquad - \lambda\left(q_n\log\frac{q_n}{q_n + q_{n+1}} + q_{n+1}\log\frac{q_{n+1}}{q_n + q_{n+1}}\right) \\ &= -\lambda\sum_{j=1}^{n+1} q_j\log q_j. \end{aligned}$$

(3) ⇒ (1). Assume that (3) is true. (1°) is shown in Theorem 2 (3), (2°) is clear and (3°) can be verified in a similar manner as in the proof of Theorem 2 (6). □