average, is greater than the lesser of the two information gains and less than the greater of the two gains, i.e.,

$$\min\{H(p_0), H(p_1)\} \leq \pi_0 H(p_0) + \pi_1 H(p_1) \leq \max\{H(p_0), H(p_1)\} \ . \tag{2.100}$$

Now consider the opposing case where the identity of the distribution to be sampled remains unknown. In this case, the most one can say about which outcome will occur in the sampling is that it is controlled by the probability distribution $p(b) = \pi_0 p_0(b) + \pi_1 p_1(b)$. In other words, the sampling outcome will be even more unpredictable than it was in either of the two individual cases; some of the unpredictability will be due to the indeterminism $p_0(b)$ and $p_1(b)$ describe and some of the unpredictability will be due to the fact that the individual distribution from which the sample is drawn remains unknown. Hence it must be the case that $H(p) \geq H(p_0)$ and $H(p) \geq H(p_1)$.

The excess of $H(p)$ over $\pi_0 H(p_0) + \pi_1 H(p_1)$ is the average gain of information one can expect about the distribution itself. This quantity, called the *mutual information* [60, 61],

$$J(p_0, p_1; \pi_0, \pi_1) = H(\pi_0 p_0 + \pi_1 p_1) - \Big(\pi_0 H(p_0) + \pi_1 H(p_1)\Big) \ , \tag{2.101}$$

is the natural candidate for distinguishability that we seek in this section. If the two distributions $p_0(b)$ and $p_1(b)$ are completely distinguishable, then all the information gained in a sampling should be solely about the identity of the distribution; the quantity $J(p_0, p_1; \pi_0, \pi_1)$ should reduce to $H(\pi)$, the information that can be gained by sampling the prior distribution $\pi = \{\pi_0, \pi_1\}$. If the distributions $p_0(b)$ and $p_1(b)$ are completely indistinguishable, then $J(p_0, p_1; \pi_0, \pi_1)$ should reduce to zero; this signifies that in sampling one learns nothing whatsoever about the distribution from which the sample is drawn.

Notice that this distinguishability measure depends crucially on the observer's prior state of knowledge, quantified by $\pi = \{\pi_0, \pi_1\}$, about whether $p_0(b)$ or $p_1(b)$ is actually the case. Thus it is a measure of distinguishability relative to a given state of knowledge. There is, of course, nothing wrong with this, just as there was nothing wrong with the error-probability distinguishability measure; one just needs to recognize it as such.

These are the ideas behind taking mutual information as a measure of distinguishability. In the remainder of this section, we work toward justifying a precise expression for Eq. (2.101) and showing in a detailed way how it can be interpreted in an operational context.

2.4.1 Derivation of Shannon's Information Function

The function $H(p)$ that quantifies the average information gained upon sampling a distribution $p(b)$ will ultimately turn out to be the famous Shannon information function [60, 62]

$$H(p) = -\sum_b p(b) \ln p(b) \ . \tag{2.102}$$

What we should like to do here is justify this expression from first principles. That is to say, we shall build up a theory of "information gain" based solely on the probabilities in an experiment and find that that theory gives rise to the expression (2.102).

To start with our most basic assumption, we reiterate the idea that the information gained in performing an experiment or observation is a function of how well the outcomes to that experiment or observation can be predicted in the first place. Other characteristics of an outcome that might convey "information" in the common sense of the word, such as shape, color, smell, feel, etc., will be considered irrelevant; indeed, we shall assume any such properties already part of the very definition of the outcome events. Formally this means that if a set of events $\{x_1, x_2, \ldots, x_n\}$ has a probability distribution $p(x)$, not only is the expected information gain in a sampling, $H(p)$,

exclusively a function of the numbers $p(x_1)$, $p(x_2)$, ..., $p(x_n)$, but also it must be independent of the labelling of that set. In other words, $H(p) \equiv H(p(x_1), p(x_2), \ldots, p(x_n))$ is required to be invariant under permutations of its arguments. This is called the requirement of "symmetry."

The most important technical property of $H(p)$ is that, even though information gain is a subjective concept depending on the observer's prior state of knowledge, it should at least be objective enough that it not depend on the method by which knowledge of the experimental outcomes is acquired. We can make this idea firm with a simple example. Consider an experiment with three mutually exclusive outcomes x, y, and z. Note that the probability that z does not occur is

$$p(\neg z) = 1 - p(z) = p(x) + p(y) \; . \tag{2.103}$$

The probabilities for x and y given that z does not occur are

$$p(x|\neg z) = \frac{p(x)}{p(x) + p(y)} \qquad \text{and} \qquad p(y|\neg z) = \frac{p(y)}{p(x) + p(y)} \; . \tag{2.104}$$

There are at least two methods by which an observer can gather the result of this experiment. The first method is by the obvious tack of simply finding which outcome of the three possible ones actually occurred. In this case, the expected information gain is, by our convention,

$$H\Big(p(x),\, p(y),\, p(z)\Big) \; . \tag{2.105}$$

The second method is more roundabout. One could, for instance, first check whether z did or did not occur, and *then* in the event that it did *not* occur, further check which of x and y *did* occur. In the first phase of this method, the expected information gain is

$$H\Big(p(\neg z),\, p(z)\Big) \; . \tag{2.106}$$

For those cases in which the second phase of the method must be carried out, a further gain of information can be expected. Namely,

$$H\Big(p(x|\neg z),\, p(y|\neg z)\Big) \; . \tag{2.107}$$

Note, though, that this last case is only expected to occur a fraction $p(\neg z)$ of the time. Thus, in total, the expected information gain by this more roundabout method is

$$H\Big(p(\neg z),\, p(z)\Big) \; + \; p(\neg z)\, H\Big(p(x|\neg z),\, p(y|\neg z)\Big) \; . \tag{2.108}$$

The assumption of "objectivity" is that the quantities in Eqs. (2.105) and (2.108) are identical. That is to say, upon changing the notation slightly to $p_x = p(x)$, $p_y = p(y)$, $p_z = p(z)$

$$H(p_x,\, p_y,\, p_z) \; = \; H(p_x + p_y \,,\, p_z) \; + \; (p_x + p_y)\, H\left(\frac{p_x}{p_x + p_y} \,,\, \frac{p_y}{p_x + p_y}\right) \; . \tag{2.109}$$

In the event that we are instead concerned with n mutually exclusive events, the same assumption of "objectivity" leads to the identification,

$$H(p_1,\, \ldots \,, p_n) \; = \; H(p_1 + p_2 \,, p_3 \,, \ldots \,, p_n) \; + \; (p_1 + p_2)\, H\left(\frac{p_1}{p_1 + p_2} \,,\, \frac{p_2}{p_1 + p_2}\right) \; . \tag{2.110}$$

It turns out that the requirements of symmetry and objectivity (as embodied in Eq. (2.110)) are enough to uniquely determine the form of $H(p)$ (up to a choice of units) provided we allow

ourselves one extra convenience [63], namely, that we allow the introduction of an arbitrary positive parameter $\alpha \neq 1$ into Eq. (2.110) in the following way,

$$H_\alpha(p_1, \ldots, p_n) \;=\; H_\alpha(p_1 + p_2\,, p_3\,, \ldots, p_n) \;+\; (p_1 + p_2)^\alpha \, H_\alpha\!\left(\frac{p_1}{p_1 + p_2}\,,\, \frac{p_2}{p_1 + p_2}\right)\,, \tag{2.111}$$

and define $H(p)$ to be the limiting value of $H_\alpha(p)$ as $\alpha \to 1$. (The introduction of the subscript on $H_\alpha(p)$ is made simply to remind us that the solutions to Eq. (2.111) depend upon the parameter α.) This idea is encapsulated in the following theorem.

Theorem 2.7 (Daróczy) *Let*

$$\Gamma_n = \left\{ (p_1, \ldots, p_n) \mid p_k \geq 0,\ k = 1, \ldots, n,\ \textit{and}\ \sum_{i=1}^{n} p_i = 1 \right\} \tag{2.112}$$

be the set of all n-point probability distributions and let $\Gamma = \bigcup_n \Gamma_n$ be the set of all discrete probability distributions. Suppose the function $H_\alpha : \Gamma \to \mathbb{R}$, $\alpha \neq 1$, is symmetric in all its arguments and satisfies Eq. (2.111) for each $n \geq 2$. Then, under the convention that $0 \ln 0 = 0$, the limiting value of H_α as $\alpha \to 1$ is uniquely specified up to a constant C by

$$H(p_1, \ldots, p_n) \;=\; -\frac{C}{\ln 2} \sum_{i=1}^{n} p_i \ln p_i\,. \tag{2.113}$$

The constant C in this expression fixes the "units" of information. If $C = 1$, information is said to be measured in *bits*; if $C = \ln 2$, information is said to be measured in *nats*. (A relatively obscure measure of information is the case $C = \log_{10} 2$, where the units are called *Hartleys* [64].) In this document, we will generally take $C = \ln 2$. On the occasion, however, that we do consider information in units of bits we shall write $\log()$ for the base-2 logarithm, rather than the more common $\log_2()$.

$\triangle$ The proof of Theorem 2.7, deserving wider recognition, is due to Daróczy [65] and proceeds as follows. Define $s_i = p_1 + \cdots + p_i$ and let

$$f(x) = H_\alpha(x, 1 - x) \qquad \text{for} \qquad 0 \leq x \leq 1\,. \tag{2.114}$$

Then, by repeated application of condition (2.111), it follows immediately that

$$H_\alpha(p_1, \ldots, p_n) \;=\; \sum_{i=2}^{n} s_i^\alpha \, f\!\left(\frac{p_i}{s_i}\right)\,. \tag{2.115}$$

Thus all we need do is focus on finding an explicit expression for the function f.

We have from the symmetry requirement that $H_\alpha(x, 1 - x) = H_\alpha(1 - x, x)$ and hence,

$$f(x) = f(1 - x)\,. \tag{2.116}$$

In particular, $f(0) = f(1)$. Furthermore, if x and y are two nonnegative numbers such that $x + y \leq 1$, we must also have

$$H_\alpha(x, y, 1 - x - y) = H_\alpha(y, x, 1 - x - y)\,. \tag{2.117}$$

However, by Eq. (2.111)

$$\begin{aligned} H_\alpha(x,y,1-x-y) &= H_\alpha(x,1-x) + (1-x)^\alpha H_\alpha\Big(\frac{y}{1-x},\frac{1-x-y}{1-x}\Big) \\ &= H_\alpha(x,1-x) + (1-x)^\alpha H_\alpha\Big(\frac{y}{1-x},1-\frac{y}{1-x}\Big) \\ &= f(x) + (1-x)^\alpha f\Big(\frac{y}{1-x}\Big) \ . \end{aligned} \tag{2.118}$$

Thus it follows that f must satisfy the functional equation

$$f(x) + (1-x)^\alpha f\Big(\frac{y}{1-x}\Big) = f(y) + (1-y)^\alpha f\Big(\frac{x}{1-y}\Big) \ , \tag{2.119}$$

for $x, y \in [0,1)$ with $x+y \leq 1$. (In the case $\alpha = 1$, Eq. (2.119) is known commonly as *the fundamental equation of information* [66].)

We base the remainder of our conclusions on the study of Eq. (2.119). Note first that if $x = 0$, it reduces to,

$$f(0) + f(y) = f(y) + (1-y)^\alpha f(0) \ . \tag{2.120}$$

Since y is still arbitrary, it follows from this that $f(0) = 0$; thus $f(1) = 0$, too. Now let $p = 1-x$ for $x \neq 1$ and let $q = y/(1-x) = y/p$. With this, the information equation (2.119) becomes

$$f(p) + p^\alpha f(q) = f(pq) + (1-pq)^\alpha f\Big(\frac{1-p}{1-pq}\Big) \ . \tag{2.121}$$

We can use this equation to show that

$$F(p,q) \equiv f(p) + \left[p^\alpha + (1-p)^\alpha\right] f(q) \tag{2.122}$$

is symmetric in q and p, i.e., $F(p,q) = F(q,p)$. From that fact, a unique expression for $f(p)$ follows trivially. Let us just show this before going further:

$$\begin{aligned} 0 &= F\Big(p\,,\,\frac{1}{2}\Big) - F\Big(\frac{1}{2}\,,\,p\Big) \\ &= f(p) + \left[p^\alpha + (1-p)^\alpha\right] f\Big(\frac{1}{2}\Big) - f\Big(\frac{1}{2}\Big) - \left[\Big(\frac{1}{2}\Big)^\alpha + \Big(\frac{1}{2}\Big)^\alpha\right] f(p) \\ &= \Big(1 - 2^{1-\alpha}\Big) f(p) + f\Big(\frac{1}{2}\Big)\left[p^\alpha + (1-p)^\alpha - 1\right] \ , \end{aligned} \tag{2.123}$$

which implies that

$$f(p) = C\left(2^{1-\alpha} - 1\right)^{-1}\left[p^\alpha + (1-p)^\alpha - 1\right] \ , \tag{2.124}$$

where the constant $C = f(\frac{1}{2})$. Because $f(0) = f(1) = 0$, Eq. (2.124) also holds for $p = 0$ and $p = 1$.

To cap off the derivation of Eq. (2.124), let us demonstrate that $F(p,q)$ is symmetric. Just expanding and regrouping, we have, by Eq. (2.121), that

$$\begin{aligned} F(p,q) &= [f(p) + p^\alpha f(q)] + (1-p)^\alpha f(q) \\ &= f(pq) + (1-pq)^\alpha f\Big(\frac{1-p}{1-pq}\Big) + (1-p)^\alpha f(q) \\ &= f(pq) + (1-pq)^\alpha\left[f\Big(\frac{1-p}{1-pq}\Big) + \Big(\frac{1-p}{1-pq}\Big)^\alpha f(q)\right] \ . \end{aligned} \tag{2.125}$$

If we can show that the last term in this expression is symmetric in q and p, then we will have shown that $F(p,q)$ is symmetric. To this end, let us define

$$A(p,q) = f\left(\frac{1-p}{1-pq}\right) + \left(\frac{1-p}{1-pq}\right)^{\alpha} f(q)\ . \tag{2.126}$$

Also, to save a little room, let

$$z = \frac{1-p}{1-pq}\ . \tag{2.127}$$

Then,

$$1 - zq = \frac{1-q}{1-pq} \qquad \text{and} \qquad 1 - z = p\left(\frac{1-q}{1-pq}\right)\ . \tag{2.128}$$

So that, upon using Eq. (2.121) again, we get

$$\begin{aligned} A(p,q) &= f(z) + z^{\alpha} f(q) \\ &= f(zq) + (1-zq)^{\alpha} f\left(\frac{1-z}{1-zq}\right) \\ &= f(1-zq) + (1-zq)^{\alpha} f\left(\frac{1-z}{1-zq}\right) \\ &= f\left(\frac{1-q}{1-pq}\right) + \left(\frac{1-q}{1-pq}\right)^{\alpha} f(p) \\ &= A(q,p)\ . \end{aligned} \tag{2.129}$$

Thus $F(p,q)$ is symmetric. This completes the demonstration of Eq. (2.124).

We just need plug the expression for $f(p)$ into Eq. (2.115) to get a nearly final result,

$$\begin{aligned} H_{\alpha}(p_1, \ldots, p_n) &= \sum_{i=2}^{n} s_i^{\alpha} C \left(2^{1-\alpha} - 1\right)^{-1} \left[\left(\frac{p_i}{s_i}\right)^{\alpha} + \left(1 - \frac{p_i}{s_i}\right)^{\alpha} - 1\right] \\ &= C \left(2^{1-\alpha} - 1\right)^{-1} \sum_{i=2}^{n} \left[\, p_i^{\alpha} + (s_i - p_i)^{\alpha} - s_i^{\alpha} \,\right] \\ &= C \left(2^{1-\alpha} - 1\right)^{-1} \sum_{i=2}^{n} \left(p_i^{\alpha} + s_{i-1}^{\alpha} - s_i^{\alpha}\right) \\ &= C \left(2^{1-\alpha} - 1\right)^{-1} \left(\sum_{i=2}^{n} p_i^{\alpha} + s_1^{\alpha} - s_n^{\alpha}\right) \\ &= C \left(2^{1-\alpha} - 1\right)^{-1} \left(\sum_{i=1}^{n} p_i^{\alpha} - 1\right)\ . \end{aligned} \tag{2.130}$$

Now in taking the limit $\alpha \to 1$, note that both the numerator and denominator of this expression vanishes. Thus we must use l'Hospital's rule in the calculating limit, i.e., first take the derivative with respect to α of the numerator and denominator separately and then take the limit:

$$\lim_{\alpha \to 0} H_{\alpha}(p_1, \ldots, p_n) = \lim_{\alpha \to 0} C \Big(-2^{1-\alpha} \ln 2\Big)^{-1} \left(\sum_{i=1}^{n} p_i^{\alpha} \ln p_i\right)$$

$$= \quad -\frac{C}{\ln 2}\sum_{i=1}^{n} p_i \ln p_i \ . \tag{2.131}$$

This completes our derivation of the Shannon information formula (2.102). It is to be hoped that this has conveyed something of the austere origin of the information-gain concept. □

We finally mention that the Daróczy informations of type-α, i.e., Eq. (2.130), appearing in this derivation are of interest in their own right. First of all, there is a simple relation between these and the Renyi informations of degree-α introduced in Section 2.2; namely,

$$\begin{aligned} H^{\alpha}(p) \quad &\equiv \quad \frac{1}{\alpha - 1}\ln\left(\sum_{i=1}^{n} p_i^{\alpha}\right) \\ &= \quad \frac{1}{\alpha - 1}\ln\left(\frac{1}{C}\Big(2^{1-\alpha} - 1\Big)H_{\alpha}(p) + 1\right) \ . \end{aligned} \tag{2.132}$$

Secondly, they share many properties with the Shannon information [66] while being slightly more tractable for some applications, there being no logarithm in their expression.

2.4.2 An Interpretation of the Shannon Information

The justification of the information-gain concept can be strengthened through an operational approach to the question. To carry this out, let us develop the following example. Suppose we were to perform an experiment with four possible outcomes x_1, x_2, x_3, x_4, the respective probabilities being $p(x_1) = \frac{1}{20}$, $p(x_2) = \frac{1}{5}$, $p(x_3) = \frac{1}{4}$, and $p(x_4) = \frac{1}{2}$. The expected gain of information in this experiment is given by Eq. (2.102) and is numerically approximately 1.68 bits. By the fundamental postulate of Section 2.4.1, we know that this information gain will be independent of the method of questioning used in discerning the outcome. In particular, we could consider all possible ways of determining the outcome by way of binary yes/no questions. For instance, we could start by asking, "Is the outcome x_1?" If the answer is yes, then we are done. If the answer is no, then we could further ask, "Is the outcome x_2?," and proceed in similar fashion until the identity of the outcome is at hand. This and three other such binary-question methodologies are depicted schematically in Figure 2.1.

The point of interest to us here is that each such questioning scheme generates, by its very nature, a *code* for the possible outcomes to the experiment. That code can be generated by writing down, in order, the yes's and no's encountered in traveling from the *root* to each *leaf* of these schematic *trees*. For instance, by substituting 0 and 1 for yes and no, respectively, the four trees depicted in Figure 2.1 give rise to the codings:

Scheme 1	Scheme 2	Scheme 3	Scheme 4
$x_1 \leftrightarrow 0$	$x_1 \leftrightarrow 00$	$x_1 \leftrightarrow 11$	$x_1 \leftrightarrow 011$
$x_2 \leftrightarrow 10$	$x_2 \leftrightarrow 01$	$x_2 \leftrightarrow 0$	$x_2 \leftrightarrow 010$
$x_3 \leftrightarrow 110$	$x_1 \leftrightarrow 10$	$x_1 \leftrightarrow 100$	$x_3 \leftrightarrow 00$
$x_4 \leftrightarrow 111$	$x_1 \leftrightarrow 11$	$x_1 \leftrightarrow 101$	$x_4 \leftrightarrow 1$

Codes that can be generated from trees in this way are called *instantaneous* or *prefix-free* and are noteworthy for the property that concatenated strings of their codewords can be uniquely deciphered just by reading from left to right. This follows because no codeword in such a coding