Chapter 2

Information and Entropy

2.1 How is information quantified?

Information theory leaped fully clothed from the forehead of Claude Shannon in 1948 [63]. The foundation of the theory is a quantification of information, a quantification that a few researchers had been floundering toward for 20 or 30 years (see [29] and [56]). The definition will appear strange and unnatural at first glance. The purpose of this first section of Chapter 2 is to acquaint the reader with certain issues regarding this definition, and finally to present the brilliant proof of its inevitability due, as far as we know, to Aczél and Daroczy in 1975 [1].

To begin to make sense of what follows, think of the familiar quantities *area* and *volume*. These are associated with certain kinds of objects—planar regions or regions on surfaces, in the case of area; bodies in space, in the case of volume. The assignment of these quantities to appropriate objects is *defined*, and the definitions can be quite involved; in fact, the final chapter on the question of mere definitions of area and volume was perhaps not written until the twentieth century, with the introduction of Lebesgue measure.

These definitions are not simply masterpieces of arbitrary mathematical cleverness—they have to respond to certain human agreements on the nature of these quantities. The more elaborate definitions have to agree with simpler ways of computing area on simple geometric figures, and a planar region composed of two other non-overlapping planar regions should have area equal to the sum of the areas of the two.

The class of objects to which the quantity *information* will be attached are occurrences of events associated with probabilistic experiments; another name for this class is *random phenomena*. It is supposed[1] that every such event or phenomenon E has a pre-assigned, a priori probability $P(E)$ of occurrence. Here is Shannon's definition of the "self-information" $I(E)$ of an event E:

$$I(E) = \log 1/P(E) = -\log P(E).$$

[1] Take care! Paradoxes and absurdities are known to be obtainable by loose manipulation of this assumption. These are avoidable by staying within the strict framework of a well-specified probabilistic experiment, in each situation.

If $P(E) = 0$, $I(E) = \infty$.

Feinstein [17] used other terminology that many would find more explanatory than "self-information": $I(E)$ is the amount of information *disclosed* or *given off* by the occurrence of E. This terminology coaxes our agreement to the premise that information ought to be a quantity attached to random phenomena with prior probabilities. And if that is agreed to, then it seems unavoidable that the quantity of information must be some function of the prior probability, i.e., $I(E) = f(P(E))$ for some function f, just because prior probability is the only quantity associated with all random phenomena, the only thing to work with.

Lest this seem a frightful simplification of the wild world of random phenomena, to compute information content as a function of prior probability alone, let us observe that this sort of simplification happens with other quantities; indeed, such simplification is one of the charms of quantification. Planar regions of radically different shapes and with very different topological properties can have the same area. Just so; why shouldn't a stock market crash in Tokyo and an Ebola virus outbreak in the Sudan possibly release the same amount of information? The quantification of information should take no account of the "quality" or category of the random phenomenon whose occurrence releases the information.

Suppose we agree that $I(E)$ ought to equal $f(P(E))$ for some function f defined on $(0,1]$, at least, for all probabilistic events E. Then why did Shannon take $f(x) = \log(1/x)$? (And which log are we talking about? But we will deal with that question in the next subsection.) We will take up the question of Shannon's inspiration, and Aczél and Daroczy's final word on the matter, in Section 2.1.3. But to get acclimated, let's notice some properties of $f(x) = \log(1/x)$, with log to any base > 1: f is a decreasing, non-negative function on $(0,1]$, and $f(1) = 0$. These seem to be necessary properties for a function to be used to quantify information, via the equation $I(E) = f(P(E))$. Since $I(E)$ is to be a quantity, it should be non-negative. The smaller the prior probability of an event the greater the quantity of information released when it occurs, so f should be decreasing. And an event of prior probability 1 should release no information at all when it occurs, so $f(1)$ should be 0.

Even among functions definable by elementary formulas, there are an infinite number of functions on $(0,1]$ satisfying the requirements noted above; for instance, $1 - x^q$ and $(1/x)^q - 1$ satisfy those requirements, for any $q > 0$. One advantage that $\log(1/x)$ has over these functions is that it converts products to sums, and a lot of products occur in the calculation of probabilities. As we shall see in section 2.1.3, this facile, shallow observation in favor of $\log(1/x)$ as the choice of function to be used to quantify information is remarkably close to the reason why $\log(1/x)$ is the only possible choice for that purpose.

2.1.1 Naming the units

For any $a, b > 0$, $a \neq 1 \neq b$, and $x > 0$, $\log_a x = (\log_a b)\log_b x$; that is, the functions $\log x$ to different bases are just constant multiples of each other. So, in Shannon's use of log in the quantification of information, changing bases is like changing units. Choosing a base amounts to choosing a unit of information. What requires discussion is the name of the unit that Shannon chose when the base is 2: Shannon chose to call that unit a *bit*.

Yes, the unit name when $\log = \log_2$ is the very same abbreviation of "binary digit" widely reported to have been invented by J. W. Tukey, who was at Bell Labs with Shannon in the several years before [63] appeared. (In [76] we read that the word "bit", with the meaning of "binary digit", first appeared *in print* in "A mathematical theory of communication.")

Now, we do not normally pay much attention to unit names in other contexts. For example, "square meter" as a unit of area seems rather self-explanatory. But in this case the connection between "bit" as a unit of information, an arbitrarily divisible quantifiable substance, like a liquid, and "bit" meaning a binary digit, either 0 or 1, is not immediately self-evident to human intuition; yet Shannon uses the two meanings interchangeably, as has virtually every other information theorist since Shannon (although Solomon Golomb, in [24], is careful to distinguish between the two). We shall attempt to justify the unit name, and, in the process, to throw light on the meaning of the information unit when the base of the logarithm is a positive integer greater than 2.

Think of one square meter of area as the greatest amount of area that can be squeezed into a square of side length one meter. (You may object that when one has a square of side length 1 meter, one already has a maximum area "squeezed" into it. Fine; just humor us on this point.) Reciprocally, the square meter measure of the area of a planar region is the side length, in meters, of the smallest square into which the region can hypothetically be squeezed, by deformation without shrinking or expanding (don't ask for a rigorous definition here!).

With this in mind, let us take, in analogy to a planar region, an entire probabilistic experiment, initially unanalyzed as to its possible outcomes; and now let it be analyzed, the possible outcomes broken into a list $E_1, \dots, E_m$ of pairwise mutually exclusive events which exhaust the possibilities: $P(\cup_{i=1}^{m} E_i) = \sum_{i=1}^{m} P(E_i) = 1$ (recall 1.2.3). If you wish, think of each E_i as a single outcome, in a set of outcomes. Assume $P(E_i) > 0, i = 1, \dots, m$.

It may be objected that rather than analogizing a planar region by an entire probabilistic experiment, a planar region to which the quantity area is assigned should be analogous to the kind of thing to which the quantity information is assigned, namely a single event. This is a valid objection.

In what follows, rather than squeezing the information contained in a single event into a "box" of agreed size, we will be squeezing the information contained in the ensemble of the events $E_1, \dots, E_m$ into a very special box, the set of binary words of a certain length. We will compare the average information content of the events $E_1,\dots,E_m$ with that length. This comparison will be taken to indicate what the maximum average (over a list like $E_1,\dots,E_m$) number of units of information can be represented by the typical (aren't they all?) binary word of that length.

We admit that this is all rather tortuous, as a justification for terminology. Until someone thinks of something better, we seem to be forced to this approach by the circumstance that we are trying to squeeze the information content of events into binary words, whereas, in the case of area, we deform a region to fit into another region of standard shape. If we considered only one event, extracted without reference to the probabilistic experiment to which it is associated, we could let it be named with a single bit, 0 or 1, and this does not seem to be telling us anything. Considering a non-exhaustive ensemble of events associated with the same probabilistic experiment (pairwise mutually exclusive so that their information contents are separate) we have a generalization of the situation with a single event; we can store a lot of information by encoding with relatively short binary words, just because we are ignoring the full universe of possibilities. Again, this does not seem to lead to a satisfactory conclusion about the relation between information and the length of binary words required to store it. What about looking at ensembles of events from possibly different probabilistic experiments? Again, unless there is some constraint on the number of these events and their probabilities, it does not seem that encoding these as binary words of fixed length tells us anything about units of information, any more than in the case when the events are associated with the same probabilistic experiment.

We realize that this discussion is not wholly convincing; perhaps someone will develop a more compelling way of justifying our setup in the future. For now, let us return to $E_1,\dots,E_m$, pairwise mutually exclusive events with $\sum_{i=1}^m P(E_i) = 1$. If we agree that $I(E) = -\log P(E)$ for any event E, with log to some base > 1, then the average information content of an event in the list $E_1,\dots,E_m$ is

$$H(E_1,\dots,E_m) = \sum_{i=1}^m P(E_i)I(E_i) = -\sum_{i=1}^m P(E_i)\log P(E_i).$$

(Recall Section 1.8.)

As is conventional, let $\ln = \log_e$, the natural logarithm.

2.1.1 Lemma *For $x > 0, \ln x \le x-1$, with equality when and only when $x = 1$.*

Indication of proof Apply elementary calculus to $f(x) = x - 1 - \ln x$ to see that $f(x) \ge 0$ on $(0,\infty)$, with equality only when $x = 1$.

2.1.2 Theorem *If $p_1,\dots,p_m$ are positive numbers summing to 1, then $-\sum_{i=1}^m p_i \log p_i \le \log m$, with equality if and only if $p_i = 1/m, i = 1,\dots,m$.*

Proof: Let $c = \log e > 0$. Since $\sum p_i = 1$,

$$\begin{aligned}(-\sum_{i=1}^{m} p_i \log p_i) - \log m &= \sum_{i=1}^{m} p_i(\log(1/p_i) - \log m) \\ &= \sum_{i=1}^{m} p_i \log(1/(mp_i)) \\ &= c\sum_{i=1}^{m} p_i \ln(1/(mp_i)) \leq c\sum_{i=1}^{m} p_i(\frac{1}{mp_i} - 1) \\ &= c\big(\sum_{i=1}^{m}(1/m) - \sum_{i=1}^{m} p_i\big) = c(1-1) = 0,\end{aligned}$$

by Lemma 2.1.1, with equality if and only if $1/(mp_i) = 1$ for each $i = 1, \ldots, m$. □

Now back to considering $E_1, \ldots, E_m$. Let k be an integer such that $m \leq 2^k$ and let us put the E_i in one-to-one correspondence with m of the binary words of length k. That is, the E_i have been encoded, or named, by members of $\{0, 1\}^k$, and thereby we consider the ensemble $E_1, \ldots, E_m$ to be stored in $\{0, 1\}^k$.

By Theorem 2.1.2, the average information content among the E_i satisfies $H(E_1, \ldots, E_m) = -\sum_{i=1}^{m} P(E_i) \log P(E_i) \leq \log m \leq \log 2^k = k$, if $\log = \log_2$; and equality can be achieved if $m = 2^k$ and $P(E_i) = 1/m$, $i = 1, \ldots, m$. That is, the greatest average number of information units per event contained in a "system of events", as we shall call them, which can be stored as k-bit binary words, is k, if the unit corresponds to $\log = \log_2$. And that, ladies and gentlemen, is why we call the unit of information a *bit* when $\log = \log_2$.

In case $\log = \log_n$, for an integer $n > 2$, we would like to call the unit of information a nit, but the term probably won't catch on. Whatever it is called, the discussion preceding can be adapted to justify the equivalence of the information unit, when $\log = \log_n$, and a single letter of an n-element alphabet.

2.1.2 Information connecting two events

Let $\log = \log_b$ for some $b > 1$, and suppose that E and F are events in the same probability space (i.e., associated with the same probabilistic experiment).

If $P(F) > 0$, the conditional information contained in E, conditional upon F, denoted $I(E \mid F)$, is

$$I(E \mid F) = -\log P(E \mid F) = -\log \frac{P(E \cap F)}{P(F)}.$$

If $P(E \cap F) = 0$ we declare $I(E \mid F) = \infty$.

The *mutual information* of (or between) E and F, denoted $I(E, F)$, is

$$I(E, F) = \log \frac{P(E \cap F)}{P(E)P(F)}, \text{ if } P(E)P(F) > 0,$$

and $I(E,F) = 0$ otherwise, i.e., if either $P(E) = 0$ or $P(F) = 0$.

If Shannon's quantification of information is agreed to, and if account is taken of the justification of the formula for conditional probability given in Section 1.3, then there should be no perplexity regarding the definition of $I(E \mid F)$. But it is a different story with $I(E,F)$. For one thing, $I(E,F)$ can be positive or negative. Indeed, if $P(E)P(F) > 0$ and $P(E \cap F) = 0$, we have no choice but to set $I(E,F) = -\infty$; also, $I(E,F)$ can take finite negative values, as well as positive ones.

We do not know of a neat justification for the term "mutual information", applied to $I(E,F)$, but quite a strong case for the terminology can be built on circumstantial evidence, so to speak. The mutual information function and the important index based on it, the mutual information between two systems of events, to be introduced in Section 2.2, behave as one would hope that indices so named would behave. As a first instance of this behavior, consider the following, the verification of which is left as an exercise.

2.1.3 Proposition $I(E,F) = 0$ *if and only if* E *and* F *are independent events.*

2.1.3 The inevitability of Shannon's quantification of information

Shannon himself provided a demonstration (in [63]) that information must be quantified as he proposed, given that it is to be a quantity attached to random phenomena, and supposing certain other fundamental premises about its behavior. His demonstration was mathematically intriguing, and certainly contributed to the shocked awe with which "A mathematical theory of communication" was received. However, after the initial astonishment at Shannon's virtuosity wears off, one notices a certain infelicity in this demonstration, arising from the abstruseness of those certain other fundamental premises referred to above. These premises are not about information directly, but about something called *entropy*, defined in Section 2.3 as the average information content of events in a system of events. [Yes, we have already seen this average in Section 2.1.1.] Defined thus, entropy can also be regarded as a function on the space of all finite probability vectors, and it is as such that certain premises—we could call them axioms—about entropy were posed by Shannon. He then showed that if entropy, defined with respect to information, is to satisfy these axioms, then information must be defined as it is.

The problem with the demonstration has to do with our assent to the axioms. This assent is supposed to arise from a prior acquaintance with the word "entropy," connoting disorder or unpredictability, in thermodynamics or the kinetic theory of gases. Even supposing an acquaintance with entropy in those contexts, there are a couple of intellectual leaps required to assent to Shannon's axioms for entropy: why should this newly defined, information-theoretic entropy carry the connotation of the older entropy, and how does this connotation translate into the specific axioms set by Shannon?

The demonstration of Feinstein [17] is of the same sort as Shannon's, with a somewhat more agreeable set of axioms, lessening the vertigo associated with the second of the intellectual leaps mentioned above. The first leap remains. Why should we assent to requirements on something called entropy just because we are calling it entropy, a word that occurs in other contexts?

Here are some requirements directly on the function f appearing in $I(E) = f(P(E))$ enunciated by Aczél and Daroczy [1]:

(i) $f(x) \geq 0$ for all $x \in (0,1]$;

(ii) $f(x) > 0$ for all $x \in (0,1)$; and

(iii) $f(pq) = f(p) + f(q)$ for all $p,q \in (0,1]$.

Requirements (i) and (ii) have been discussed in section 2.1.1. Notice that there is no requirement that f be decreasing here.

Obviously requirement (iii) above is a strong requirement, and deserves considerable comment. Suppose that $p,q \in (0,1]$. Suppose that we can find events E and F, possibly associated with different probabilistic experiments, such that $P(E) = p$ and $P(F) = q$. (We pass over the question of whether or not probabilistic experiments providing events of arbitrary prior probability can be found.) Now imagine the two-stage experiment consisting of performing copies of the experiments associated with E and F independently. Let G be the event "E occurred in the one experiment and F occurred in the other". Then, as we know from section 1.3, $P(G) = pq$, so $I(G) = f(pq)$.

On the other hand, the independence of the performance of the probabilistic experiments means that the information given off by the occurrence of E in one, and F in the other, ought to be the sum of the information quantities disclosed by the occurrence of each separately. This is like saying that the area of a region made up of two non-overlapping regions ought to be the sum of the areas of the constituent regions. Thus we should have

$$f(pq) = I(G) = I(E) + I(F) = f(p) + f(q).$$

We leave it to the reader to scrutinize the heart of the matter, the contention that because the two probabilistic experiments are performed with indifference, or obliviousness, to each other, the information that an observer will obtain from the occurrences of E and F, in the different experiments, ought to be $I(E) + I(F)$. We make the obvious remark that if you receive something—say, money—from one source, and then some more money from a totally different source, then the total amount of money received will be the sum of the two amounts received.

We achieve the purpose of this subsection by proving a slightly stronger version of Aczél and Daroczy's result, that if f satisfies (i), (ii), and (iii), above, then $f(x) = -\log_b x$ for some $b > 1$ for all $x \in (0,1]$.

2.1.4 Theorem *Suppose that f is a real-valued function on $[0,1)$ satisfying*

(a) $f(x) \geq 0$ for all $x \in (0,1]$;

(b) $f(\alpha) > 0$ *for some* $\alpha \in (0,1]$*; and*

(c) $f(pq) = f(p) + f(q)$ *for all* $p, q \in (0,1]$.

Then for some $b > 1$, $f(x) = -\log_b x$ *for all* $x \in (0,1]$.

Proof: First we show that f is monotone non-increasing on $(0,1]$. Suppose that $0 < x < y \le 1$. Then $f(x) = f(y\frac{x}{y}) = f(y) + f(x/y) \ge f(y)$, by (a) and (c).

Now we use a standard argument using (c) alone to show that $f(x^r) = rf(x)$ for any $x \in (0,1]$ and any positive rational r. First, using (c) repeatedly, or, if you prefer, proceeding by induction on m, it is straightforward to see that for each $x \in (0,1]$ and each positive integer m, $f(x^m) = mf(x)$. Now suppose that $x \in (0,1]$ and that m and n are positive integers. Then $x^m, x^{m/n} \in (0,1]$, and

$$mf(x) = f(x^m) = f((x^{m/n})^n) = nf(x^{m/n}),$$

so $f(x^{m/n}) = \frac{m}{n} f(x)$.

By (c), $f(1) = f(1) + f(1)$, so $f(1) = 0$. Therefore the α mentioned in (b) is not 1. As b ranges over $(1,\infty)$, $-\log_b \alpha = -\frac{\ln\alpha}{\ln b}$ ranges over $(0,\infty)$. Therefore, for some $b > 1$, $-\log_b \alpha = f(\alpha)$. By the result of the paragraph preceding and the properties of $\log_b$, the functions f and $-\log_b$ agree at each point α^r, r a positive rational.

The set of such points is dense in $(0,1]$. An easy way to see this is to note that $\ln\alpha^r = r\ln\alpha$, so $\{\ln\alpha^r; r \text{ is a positive rational}\}$ is dense in $(-\infty,0)$, by the well-known density of the rationals in the real numbers; and the inverse of ln, the exponential function, being continuous and increasing, will map a dense set in $(-\infty,0)$ onto a dense set in $(0,1]$.

We have that f and $-\log_b$ are both non-increasing, they agree on a dense subset of $(0,1]$, and $-\log_b$ is continuous. We conclude that $f = -\log_b$ on $(0,1]$. The argument forcing this conclusion is left as an exercise. □

Exercises 2.1

1. Regarding the experiment described in Exercise 1.3.5, let E_A = "urn A was chosen" and F_g = "a green ball was drawn". Write explicitly: (a) $I(E_A, F_g)$; (b) $I(E_A \mid F_g)$; (c) $I(F_g \mid E_A)$.

2. Suppose that E is an event in some probability space, and $P(E) > 0$. Show that $I(E,E) = I(E)$, and that $I(E \mid E) = 0$.

3. Suppose that E and F are events in some probability space. Show that $I(E,F) = 0$ if and only if E and F are independent. Show that $I(E,F) = I(E) - I(E \mid F)$, if $P(E)P(F) > 0$. Show that $I(E,F) \le \min(I(E), I(F))$. Show that $I(E,F) = I(F)$ if and only if E is essentially contained in F, meaning, $P(E \setminus F) = 0$.

4. Fill in the proof of Lemma 2.1.1.

5. Suppose that $p_1, \dots, p_n$, $q_1, \dots, q_n$ are positive numbers and $\sum_i p_i = 1 = \sum_i q_i$. Show that $\sum_{i=1}^m p_i \log(1/q_i) \le \sum_{i=1}^m p_i \log(1/p_i)$ with equality if and only if $p_i = q_i, i = 1, \dots, n$. [Hint: look at the proof of Theorem 2.1.2.]

6. Show that if f and g are monotone non-increasing real-valued functions on a real interval I which agree on a dense subset of I, and g is continuous, then $f = g$ on I. Give an example to show that the conclusion is not valid if the assumption that g is continuous is omitted.

2.2 Systems of events and mutual information

Suppose that $(\mathcal{S}, P)$ is a finite probability space. A *system of events* in $(\mathcal{S}, P)$ is a finite indexed collection $\mathcal{E} = [E_i; i \in I]$ of pairwise mutually exclusive events satisfying $1 = P(\bigcup_{i \in I} E_i)$.

Remarks

2.2.1 When $1 = P(\bigcup_{i \in I} E_i)$, it is common to say that the E_i *exhaust* $\mathcal{S}$.

2.2.2 Note that if the E_i are pairwise mutually exclusive, then $P(\bigcup_{i \in I} E_i) = \sum_{i \in I} P(E_i)$, by 1.2.3.

2.2.3 Any *partition* of $\mathcal{S}$ is a system of events in $(\mathcal{S}, P)$ (see exercise 1.1.2), and partitioning is the most obvious way of obtaining systems of events. For instance, in the case of n Bernoulli trials, with $\mathcal{S} = \{S, F\}^n$, if we take $E_k =$ "exactly k successes," then $E_0, \dots, E_n$ partition $\mathcal{S}$.

It is possible to have a system of events in $(\mathcal{S}, P)$ which does not partition $\mathcal{S}$ only when $\mathcal{S}$ contains outcomes with zero probability. Just as it is convenient to allow outcomes of zero probability to be elements of sets of outcomes, it is convenient to allow events of zero probability in systems of events. One aspect of this convenience is that when we derive new systems from old, as we shall, we do not have to stop and weed out the events of probability zero in the resultant system.

In deriving or describing systems of events we may have repeated events in the system, $E_i = E_j$ for some indices $i \ne j$. In this case, $P(E_i) = 0$. For better or for worse, the formality of defining a system of events as an indexed collection, rather than as a set or list of events, permits such repetition.

2.2.4 If $\mathcal{E} = [E_i; i \in I]$ is a system of events in $(\mathcal{S}, P)$, we can take $\mathcal{E}$ as a new set of outcomes. The technical niceties are satisfied since $1 = P(\bigcup_{i \in I} E_i) = \sum_{i \in I} P(E_i)$; in viewing $\mathcal{E}$ as a *set* we regard the E_i as distinct, even when they are not.

Taking $\mathcal{E}$ as a new set of outcomes involves a certain change of view. The old outcomes are merged into "larger" conglomerate outcomes, the events E_i. The changes in point of view achievable in this way are constrained by the choice of the original set of outcomes, $\mathcal{S}$. Notice that $\mathcal{S}$ itself can be thought of as a system of events, if we identify each $s \in \mathcal{S}$ with the event $\{s\}$.

If $\mathcal{F}$ is a system of events in $(\mathcal{S}, P)$, and we choose to view $\mathcal{F}$ as a set of outcomes, then we can form systems of events in the "new" space $(\mathcal{F}, P)$ just as $\mathcal{F}$ is formed from $\mathcal{S}$. Systems of events in $(\mathcal{F}, P)$ are really just systems in $(\mathcal{S}, P)$ that bear a certain relation to $\mathcal{F}$.

2.2.5 Definition Suppose $\mathcal{E} = [E_i; i \in I]$ and $\mathcal{F} = [F_j; j \in J]$ are systems of events in a finite probability space $(\mathcal{S}, P)$; we say that $\mathcal{E}$ is an *amalgamation* of $\mathcal{F}$ in case for each $j \in J$ there is some $i \in I$ such that $P(E_i \cap F_j) = P(F_j)$.

2.2.6 Definition If $(\mathcal{S}, P)$ is a finite probability space and $E, F \subseteq \mathcal{S}$, we will say that $F \subseteq E$ *essentially* if $P(F \setminus E) = 0$, and $F = E$ essentially if $F \subseteq E$ essentially and $E \subseteq F$ essentially.

To make sense of Definition 2.2.5, notice that $P(E_i \cap F_j) = P(F_j)$ is equivalent to $P(F_j \setminus (E_i \cap F_j)) = 0$, which means that F_j is contained in E_i, except, possibly, for outcomes of zero probability. So, the condition in the definition says that each F_j is essentially contained in some E_i. By Corollary 2.2.8, below, $P(F_j) = \sum_{i \in I} P(E_i \cap F_j)$ for each $j \in J$, so if $0 < P(F_j) = P(E_{i_0} \cap F_j)$ for some $i_0 \in I$, then i_0 is unique and $P(E_i \cap F_j) = 0$ for all $i \in I, i \neq i_0$. This says that, when $\mathcal{E}$ is an amalgamation of $\mathcal{F}$, each F_j of positive probability is essentially contained in exactly one E_i. Since the F_j are mutually exclusive and essentially cover $\mathcal{S}$, it also follows that each E_i is essentially (neglecting outcomes of zero probability) the union of the F_j it essentially contains; i.e., the E_i are obtained by "amalgamating" the F_j somehow. (Recall exercise 1.1.2.)

Indeed, the most straightforward way to obtain amalgamations of $\mathcal{F}$ is as follows: partition J into non-empty subsets $J_1, \dots, J_k$, and set $\mathcal{E} = [E_1, \dots, E_k]$, with $E_i = \bigcup_{j \in J_i} F_j$. It is left to you to verify that $\mathcal{E}$ thus obtained is an amalgamation of $\mathcal{F}$. We shall prove the insinuations of the preceding paragraph, and see that every amalgamation is essentially (neglecting outcomes of probability zero) obtained in this way. This formality will also justify the interpretation of an amalgamation of $\mathcal{F}$ as a system of events in the new space $(\mathcal{F}, P)$, treating $\mathcal{F}$ as a set of outcomes. Readers who already see this interpretation and abhor formalities may skip to 2.2.10.

2.2.7 Lemma *If $(\mathcal{S}, P)$ is a finite probability space, $E, F \subseteq \mathcal{S}$, and $P(F) = 1$, then $P(E) = P(E \cap F)$.*

Proof: $P(E) = P(E \cap F) + P(E \cap (\mathcal{S} \setminus F)) = P(E \cap F) + 0$, since $E \cap (\mathcal{S} \setminus F) \subseteq \mathcal{S} \setminus F$ and $P(\mathcal{S} \setminus F) = 1 - P(F) = 1 - 1 = 0$. □

2.2.8 Corollary *If $\mathcal{F} = [F_j; j \in J]$ is a system of events in $(\mathcal{S}, P)$, and $E \subseteq \mathcal{S}$, then $P(E) = \sum_{j \in J} P(E \cap F_j)$.*

Proof: Suppose $j_1, j_2 \in J$ and $j_1 \neq j_2$. We have $(E \cap F_{j_1}) \cap (E \cap F_{j_2}) \subseteq F_{j_1} \cap F_{j_2}$, so $0 \leq P((E \cap F_{j_1}) \cap (E \cap F_{j_2})) \leq P(F_{j_1} \cap F_{j_2}) = 0$ since F_{j_1} and F_{j_2} are mutually exclusive. Thus $E \cap F_{j_1}$ and $E \cap F_{j_2}$ are mutually exclusive. It follows that $\sum_{j \in J} P(E \cap F_j) = P(\bigcup_{j \in J}(E \cap F_j)) = P(E \cap (\bigcup_{j \in J} F_j)) = P(E)$ by 2.2.7, taking $F = \bigcup_{j \in J} F_j$. □

2.2.9 Theorem *Suppose that $\mathcal{F} = [F_j;\, j \in J]$ is a system of events in a finite probability space $(\mathcal{S}, P)$, and I is a finite non-empty set. An indexed collection $\mathcal{E} = [E_i;\, i \in I]$ of subsets of $\mathcal{S}$ is an amalgamation of $\mathcal{F}$ if and only if there is a partition $[J_i;\, i \in I]$ of J (into not necessarily non-empty sets) such that, for each $i \in I$, $E_i = \bigcup_{j \in J_i} F_j$ essentially.*

Proof: The proof of the "if" assertion is left to the reader. Note that as part of this proof it should be verified that $\mathcal{E}$ is a system of events.

Suppose that $\mathcal{E}$ is an amalgamation of $\mathcal{F}$. If $i \in I$ and $P(E_i) > 0$, set $J_i = \{j \in J;\, P(E_i \cap F_j) > 0\}$. If $j \in J$ and $P(F_j) > 0$, then there is a unique $i_0 \in I$ such that $j \in J_{i_0}$, by the argument in the paragraph following Definition 2.2.6. Note that $P(F_j) = P(F_j \cap E_i)$ if $j \in J_i$, by that argument.

Thus we have pairwise disjoint sets $J_i \subseteq J$ containing every $j \in J$ such that $P(F_j) > 0$. If $P(F_j) = 0$, put j into one of the J_i, it doesn't matter which. If $P(E_i) = 0$, set $J_i = \emptyset$. The J_i, $i \in I$, now partition J. It remains to be seen that $E_i = \bigcup_{j \in J_i} F_j$ essentially for each $i \in I$. This is clear if $P(E_i) = 0$. If $P(E_i) > 0$, then, since $P(E_i \cap F_j) > 0$ only for $j \in J_i$, we have

$$\begin{aligned} P(E_i) &= \sum_{j \in J} P(E_i \cap F_j) \quad \text{[by Corollary 2.2.8]} \\ &= \sum_{j \in J_i} P(E_i \cap F_j) = P\Big(\bigcup_{j \in J_i} (E_i \cap F_j)\Big), \text{ so} \\ P\Big(E_i \setminus \bigcup_{j \in J_i} F_j\Big) &= P\Big(E_i \setminus \bigcup_{j \in J_i} (E_i \cap F_j)\Big) \\ &= P(E_i) - P\Big(\bigcup_{j \in J_i} (E_i \cap F_j)\Big) = 0. \end{aligned}$$

On the other hand,

$$P(E_i) = \sum_{j \in J_i} P(E_i \cap F_j) = \sum_{j \in J_i} P(F_j)$$

implies that

$$\begin{aligned} P\Big(\bigcup_{j \in J_i} F_j \setminus E_i\Big) &= P\Big(\bigcup_{j \in J_i} (F_j \setminus (E_i \cap F_j))\Big) \\ &= \sum_{j \in J_i} [P(F_j) - P(E_i \cap F_j)] = 0. \end{aligned}$$

Thus $E_i = \bigcup_{j \in J_i} F_j$, essentially. □

When $\mathcal{E}$ is an amalgamation of $\mathcal{F}$, we say that $\mathcal{E}$ is *coarser* than $\mathcal{F}$, and $\mathcal{F}$ is *finer* than $\mathcal{E}$. Given $\mathcal{E}$ and $\mathcal{F}$, neither necessarily coarser than the other, there is an obvious way to obtain a coarsest system of events which is finer than each of $\mathcal{E}$, $\mathcal{F}$. (See Exercise 2.2.2.)

2.2.10 Definition Suppose that $\mathcal{E} = [E_i; i \in I]$ and $\mathcal{F} = [F_j; j \in J]$ are systems of events in a finite probability space $(\mathcal{S}, P)$. The *joint system* associated with $\mathcal{E}$ and $\mathcal{F}$, denoted $\mathcal{E} \wedge \mathcal{F}$, is

$$\mathcal{E} \wedge \mathcal{F} = [E_i \cap F_j; (i, j) \in I \times J].$$

$\mathcal{E} \wedge \mathcal{F}$ is also called the *join* of $\mathcal{E}$ and $\mathcal{F}$.

2.2.11 Theorem *If $\mathcal{E}$ and $\mathcal{F}$ are systems of events in $(\mathcal{S}, P)$ then $\mathcal{E} \wedge \mathcal{F}$ is a system of events in $(\mathcal{S}, P)$.*

Proof: If $(i, j)_, (i', j') \in I \times J$ and $(i, j) \neq (i', j')$, then either $i \neq i'$ or $j \neq j'$. Suppose that $i \neq i'$. Since $(E_i \cap F_j) \cap (E'_i \cap F'_j) \subseteq E_i \cap E'_i$, we have

$$0 \leq P((E_i \cap F_j) \cap (E'_i \cap F'_j)) \leq P(E_i \cap E'_i) = 0,$$

so $E_i \cap F_j$ and $E'_i \cap F'_j$ are mutually exclusive. The case $i = i'$ but $j \neq j'$ is handled symmetrically.

Next, since mutual exclusivity has already been established,

$$\begin{aligned} P\Big(\bigcup_{(i,j) \in I \times J} (E_i \cap F_j)\Big) &= \sum_{(i,j) \in I \times J} P(E_i \cap F_j) \\ &= \sum_{i \in I} \sum_{j \in J} P(E_i \cap F_j) \\ &= \sum_{i \in I} P(E_i) \qquad \text{[by Corollary 2.2.8]} \\ &= 1. \end{aligned}$$

□

Definition Suppose that $\mathcal{E} = [E_i; i \in I]$ and $\mathcal{F} = [F_j; j \in J]$ are systems of events in some finite probability space $(\mathcal{S}, P)$; $\mathcal{E}$ and $\mathcal{F}$ are *statistically independent* if and only if E_i and F_j are independent events, for each $i \in I$ and $j \in J$.

Statistically independent systems of events occur quite commonly in association with multistage experiments in which two of the stages are "independent" – i.e., outcomes at one of the two stages do not influence the probabilities of the outcomes at the other. For instance, think of an experiment consisting of flipping a coin, and then drawing a ball from an urn, and two systems, one consisting of the two events "heads came up", "tails came up", with reference to the coin flip, and the other consisting of the events associated with the colors of the balls that might be drawn. For a more formal discussion of stage, or component, systems associated with multistage experiments, see Exercise 2.2.5.

2.2.12 Definition Suppose that $\mathcal{E} = [E_i; i \in I]$ and $\mathcal{F} = [F_j; j \in J]$ are systems of events in a finite probability space $(\mathcal{S}, P)$. The *mutual information between* $\mathcal{E}$ and $\mathcal{F}$ is

$$I(\mathcal{E}, \mathcal{F}) = \sum_{i \in I} \sum_{j \in J} P(E_i \cap F_j) I(E_i, F_j).$$

In the expression for $I(\mathcal{E},\mathcal{F})$ above, $I(E_i,F_j)$ is the mutual information between events E_i and F_j as defined in Section 2.1.2: $I(E_i,F_j) = \log\frac{P(E_i\cap F_j)}{P(E_i)P(F_j)}$ if $P(E_i)P(F_j) > 0$, and $I(E_i,F_j) = 0$ otherwise. If we adopt the convention that $0\log(\text{anything}) = 0$, then we are permitted to write

$$I(\mathcal{E},\mathcal{F}) = \sum_{i\in I}\sum_{j\in J} P(E_i\cap F_j)\log\frac{P(E_i\cap F_j)}{P(E_i)P(F_j)};$$

this rewriting will turn out to be a great convenience.

The mutual information between two systems of events will be an extremely important parameter in the assessment of the performance of communication channels, in Chapters 3 and 4, so it behooves us to seek some justification of the term "mutual information." Given $\mathcal{E}$ and $\mathcal{F}$, we can think of the mapping $(i,j)\to I(E_i,F_j)$ as a random variable on the system $\mathcal{E}\wedge\mathcal{F}$, and we then see that $I(\mathcal{E},\mathcal{F})$ is the average value of this random variable, the "mutual information between events" random variable. This observation would justify the naming of $I(\mathcal{E},\mathcal{F})$, if only we were quite sure that the mutual-information-between-events function is well named. It seems that the naming of both mutual informations, between events and between systems of events, will have to be justified by the behavior of these quantities. We have seen one such justification in Proposition 2.1.3, and there is another in the next Theorem. This theorem, by the way, is quite surprising in view of the fact that the terms in the sum defining $I(\mathcal{E},\mathcal{F})$ can be negative.

2.2.13 Theorem *Suppose that $\mathcal{E}$ and $\mathcal{F}$ are systems of events in a finite probability space. Then $I(\mathcal{E},\mathcal{F}) \geq 0$ with equality if and only if $\mathcal{E}$ and $\mathcal{F}$ are statistically independent.*

Proof: Let $c = \log e$, so that $\log x = c\ln x$ for all $x > 0$. If $i\in I$, $j\in J$, and $P(E_i\cap F_j) > 0$, we have, by Lemma 2.1.1,

$$\begin{aligned} P(E_i\cap F_j)\log\frac{P(E_i)P(F_j)}{P(E_i\cap F_j)} &\leq cP(E_i\cap F_j)\left[\frac{P(E_i)P(F_j)}{P(E_i\cap F_j)} - 1\right] \\ &= c\left[P(E_i)P(F_j) - P(E_i\cap F_j)\right] \end{aligned}$$

with equality if and only if $\frac{P(E_i)P(F_j)}{P(E_i\cap F_j)} = 1$, i.e., if and only if E_i and F_j are independent events. If $P(E_i\cap F_j) = 0$ then, using the convention that $0\log(\text{anything}) = 0$, we have

$$0 = P(E_i\cap F_j)\log\frac{P(E_i)P(F_j)}{P(E_i\cap F_j)} \leq c[P(E_i)P(F_j) - P(E_i\cap F_j)],$$

again, with equality if and only if $P(E_i)P(F_j) = 0 = P(E_i\cap F_j)$. Thus

$$\begin{aligned}
-I(\mathcal{E},\mathcal{F}) &= \sum_{i\in I}\sum_{j\in J} P(E_i\cap F_j)\log\frac{P(E_i)P(F_j)}{P(E_i\cap F_j)}\\
&\le c\sum_{i\in I}\sum_{j\in J}[P(E_i)P(F_j)-P(E_i\cap F_j)]\\
&= c\left[\sum_{i\in I}P(E_i)\sum_{j\in J}P(F_j)-\sum_{i\in I}\sum_{j\in J}P(E_i\cap F_j)\right]\\
&= c[1-1]=0 \qquad\qquad \text{[note Theorem 2.2.11]}
\end{aligned}$$

with equality if and only if E_i and F_j are independent, for each $i\in I$, $j\in J$. □

Exercises 2.2

1. You have two fair dice, one red, one green. You roll them once. We can make a probability space referring to this experiment in a number of different ways. Let

   $\mathcal{S}_1 =$ {"i appeared on the red die, j on the green" ; $i, j\in\{1,\dots,6\}$},
   $\mathcal{S}_2 =$ {"i appeared on one of the dice, and j on the other" ;
   $i, j\in\{1,\dots,6\}$},
   $\mathcal{S}_3 =$ {"the sum of the numbers appearing on the dice was k";
   $k\in\{2,\dots,12\}$}, and
   $\mathcal{S}_4 =$ {"even numbers on both dice", "even on the red, odd on the green",
   "even on the green, odd on the red", "odd numbers on both dice"}.

   Which pairs of these sets of outcomes have the property that neither is an amalgamation of the other?

2. Suppose that $\mathcal{E}$ and $\mathcal{F}$ are systems of events in a finite probability space.

   (a) Prove that each of $\mathcal{E},\mathcal{F}$ is an amalgamation of $\mathcal{E}\wedge\mathcal{F}$. [Thus, $\mathcal{E}\wedge\mathcal{F}$ is finer than each of $\mathcal{E},\mathcal{F}$.]

   (b) Suppose that each of $\mathcal{E},\mathcal{F}$ is an amalgamation of a system of events $\mathcal{G}$. Show that $\mathcal{E}\wedge\mathcal{F}$ is an amalgamation of $\mathcal{G}$. [So $\mathcal{E}\wedge\mathcal{F}$ is the coarsest system of events, among those that are finer than each of $\mathcal{E},\mathcal{F}$.] Here is a hint for (b): Suppose that E, F, and G are events in $\mathcal{E},\mathcal{F}$ and $\mathcal{G}$, respectively, and $P(G\cap E\cap F)>0$. Then $P(G\cap E), P(G\cap F)>0$. By the assumption that $\mathcal{E},\mathcal{F}$ are amalgamations of $\mathcal{G}$ and an argument in the proof of Theorem 2.2.9, it follows that G is essentially contained in E, and in F. So ...

3. Two fair dice, one red, one green, are rolled once. Let $\mathcal{E}=[E_1,\dots,E_6]$, where $E_i=$ "i came up on the red die", and $\mathcal{F}=[F_2,\dots,F_{12}]$, where $F_j=$ "the sum of the numbers that came up on the two dice was j". Write $I(\mathcal{E},\mathcal{F})$ explicitly, in a form that permits calculation once a base for "log" is specified.

4. Regarding the experiment described in Exercise 1.3.5, let $E_U =$ "urn U was chosen", $U \in \{A, B, C\}$, $F_r =$ "a red ball was drawn", $F_g =$ "a green ball was drawn", $\mathcal{E} = [E_A, E_B, E_C]$, and $\mathcal{F} = [F_r, F_g]$. Write $I(\mathcal{E}, \mathcal{F})$ explicitly, in a form that permits calculation, given a base for "log".

5. Suppose we have a k-stage experiment, with possible outcomes $x_1^{(i)}, \dots, x_{n_i}^{(i)}$ at the ith stage, $i = 1, \dots, k$. Let us take, as we often do, $\mathcal{S} = \{(x_{j_1}^{(i)}, \dots, x_{j_k}^{(k)}),\ 1 \le j_i \le n_i, i = 1, \dots, k\}$, the set of all sequences of possible outcomes at the different stages.

   For $1 \le i \le k$, $1 \le j \le n_i$, let $E_j^{(i)} =$ "$x_j^{(i)}$ occurred at the ith stage", and $\mathcal{E}^{(i)} = [E_j^{(i)}; j = 1, \dots, n_i]$. We will call $\mathcal{E}^{(i)}$ the ith *stage*, or ith *component*, system of events in $(\mathcal{S}, P)$. Two stages will be called *independent* if and only if their corresponding component systems are statistically independent.

   In the case $k = 2$, let us simplify notation by letting the possible outcomes at the first stage be $x_1, \dots, x_n$ and at the second stage $y_1, \dots, y_m$. Let $E_i =$ "x_i occurred at the first stage", $i = 1, \dots, n$, $F_j =$ "y_j occurred at the second stage", $j = 1, \dots, m$, be the events comprising the component systems $\mathcal{E}$ and $\mathcal{F}$, respectively.

   (a) Let $p_i = P(E_i)$ and $q_{ij} = P(F_j \mid E_i)$, $i = 1, \dots, n$, $j = 1, \dots, m$, as in Section 1.3. Assume that each p_i is positive. Show that $\mathcal{E}$ and $\mathcal{F}$ are statistically independent if and only if the q_{ij} depend only on j, not i. (That is, for any $i_1, i_2 \in \{1, \dots, n\}$ and $j \in \{1, \dots, m\}$, $q_{i_1 j} = q_{i_2 j}$.)

   (b) Verify that if $\mathcal{S}$ is regarded as a system of events in the space $(\mathcal{S}, P)$ (i.e., identify each pair (x_i, y_j) with the event $\{(x_i, y_j)\}$), then $\mathcal{S} = \mathcal{E} \wedge \mathcal{F}$.

   *(c) Suppose that three component systems of a multistage experiment, say $\mathcal{E}^{(1)}, \mathcal{E}^{(2)}$, and $\mathcal{E}^{(3)}$, are pairwise statistically independent. Does it follow that they are *jointly* statistically independent? This would mean that for all $1 \le i \le n_1$, $1 \le j \le n_2$, and $1 \le k \le n_3$, $P(E_i^{(1)} \cap E_j^{(2)} \cap E_k^{(3)}) = P(E_i^{(1)})P(E_j^{(2)})P(E_k^{(3)})$. Take note of Exercise 1.4.7.

6. (a) Suppose that $E_1, \dots, E_k, F_1, \dots, F_r$ are events in a finite probability space $(\mathcal{S}, P)$ satisfying

   (i) $E_1, \dots, E_k$ are pairwise mutually exclusive;
   (ii) $F_1, \dots, F_r$ are pairwise mutually exclusive; and
   (iii) for each $i \in \{1, \dots, k\}$, $j \in \{1, \dots, r\}$, E_i and F_j are independent.

   Show that $\bigcup_{i=1}^{k} E_i$ and $\bigcup_{j=1}^{r} F_j$ are independent.

   (b) Suppose that $\mathcal{E}$, $\widehat{\mathcal{E}}$, $\mathcal{F}$, and $\widehat{\mathcal{F}}$ are systems of events in some finite probability space, and $\widehat{\mathcal{E}}$ and $\widehat{\mathcal{F}}$ are amalgamations of $\mathcal{E}$ and $\mathcal{F}$, respectively. Show that if $\mathcal{E}$ and $\mathcal{F}$ are statistically independent, then so are $\widehat{\mathcal{E}}$ and $\widehat{\mathcal{F}}$. [You did the hard work in part (a).]

(c) It is asserted at the end of Section 1.4 that "when two stages [of a multistage experiment] are independent, any two events whose descriptions involve only those stages, respectively, will be independent." Explain how this assertion is a special case of the result of 6(a), above.

7. Succinctly characterize the systems of events $\mathcal{E}$ such that $I(\mathcal{E},\mathcal{E}) = 0$.

8. Suppose that $\mathcal{E}$ and $\mathcal{F}$ are systems of events in some finite probability space, and that $\widehat{\mathcal{E}}$ is an amalgamation of $\mathcal{E}$. Show that $I(\widehat{\mathcal{E}},\mathcal{F}) \le I(\mathcal{E},\mathcal{F})$. [You may have to use Lemma 2.1.1.]

2.3 Entropy

Suppose that $\mathcal{E} = [E_i; i \in I]$ and $\mathcal{F} = [F_j; j \in J]$ are systems of events in some finite probability space $(\mathcal{S}, P)$. The *entropy* of $\mathcal{E}$, denoted $H(\mathcal{E})$, is

$$H(\mathcal{E}) = -\sum_{i\in I} P(E_i)\log P(E_i).$$

The *joint entropy of the systems* $\mathcal{E}$ and $\mathcal{F}$ is the entropy of the joint system,

$$H(\mathcal{E}\wedge\mathcal{F}) = -\sum_{i\in I}\sum_{j\in J} P(E_i\cap F_j)\log P(E_i\cap F_j).$$

The *conditional entropy of* $\mathcal{E}$*, conditional upon* $\mathcal{F}$, is

$$\begin{aligned} H(\mathcal{E}\mid\mathcal{F}) &= \sum_{i\in I}\sum_{j\in J} P(E_i\cap F_j) I(E_i \mid F_j) \\ &= -\sum_{i\in I}\sum_{j\in J} P(E_i\cap F_j)\log\frac{P(E_i\cap F_j)}{P(F_j)}. \end{aligned}$$

Remarks

2.3.1 In the definitions above, we continue to observe the convention that $0\log(\text{anything}) = 0$. As in the preceding section, the base of the logarithm is unspecified; any base greater than 1 may be used.

2.3.2 Taking $\mathcal{E}$ as a new set of outcomes for the probability space, we see that $H(\mathcal{E})$ is the average value of the self-information of the events (now outcomes) in $\mathcal{E}$. Similarly, the joint and conditional entropies are average values of certain kinds of information.

2.3.3 The word *entropy* connotes disorder, or uncertainty, and the number $H(\mathcal{E})$ is to be taken as a measure of the disorder or uncertainty inherent in the system $\mathcal{E}$. What sort of disorder or uncertainty is associable with a system $\mathcal{E}$, and why is $H(\mathcal{E})$, as defined here, a good measure of it?

A system of events represents a way of looking at an experiment. The pesky little individual outcomes are grouped into events according to some organizing principle. We have the vague intuition that the more finely we divide the set of outcomes into events, the greater the "complexity" of our point of view, and the less order and simplicity we have brought to the analysis of the experiment. Thus, there is an intuitive feeling that some systems of events are more complex, less simple, than others, and it is not too great a stretch to make complexity a synonym, in this context, of disorder or uncertainty.

As to why H is a suitable measure of this felt complexity: as with mutual information, the justification resides in the behavior of the quantity. We refer to the theorem below, and to the result of Exercise 4 at the end of this section. The theorem says that $H(\mathcal{E})$ is a minimum (zero) when and only when $\mathcal{E}$ consists of one big event that is certain to occur, together, possibly, with massless events (events with probability zero). Surely this is the situation of maximum simplicity, minimum disorder. (A colleague has facetiously suggested that such systems of events be called "Mussolini systems". The reference is to the level of order in the system; in correlation to this terminology, an event of probability one may be called a Mussolini.) The theorem also says that, for a fixed value of $|\mathcal{E}| = |I|$, the greatest value $H(\mathcal{E})$ can take is achieved when and only when the events in $\mathcal{E}$ are equally likely. This taxes the intuition a bit, but it does seem that having a particular number of equiprobable events is a more "uncertain" or "complex" situation than having the same number of events, but with some events more likely than others.

The result of Exercise 2.3.4 is that if $\widehat{\mathcal{E}}$ is obtained from $\mathcal{E}$ by amalgamation, then $H(\widehat{\mathcal{E}}) \le H(\mathcal{E})$. To put this the other way around, if you obtain a new system from an old system by dividing the old events into smaller events, the entropy goes up, as it should.

Shannon ([63] and [65]) introduced an axiom system for entropy, a series of statements that the symbol H *ought* to satisfy to be worthy of the name *entropy*, and showed that the definition of entropy given here is the only one compatible with these requirements. As previously mentioned, Feinstein [17] did something similar, with (perhaps) a more congenial axiom system. For an excellent explanation of these axioms and further references on the matter, see the book by Dominic Welsh [81] or that of D. S. Jones [37]. We shall not pursue further the question of the validity of the definition of H, nor its uniqueness.

2.3.4 Theorem *Suppose that* $\mathcal{E} = [E_i; i \in I]$ *is a system of events in a finite probability space. Then* $0 \le H(\mathcal{E}) \le \log |I|$. *Equality at the lower extreme occurs if and only if all but one of the events in* $\mathcal{E}$ *have probability zero.* [*That one event would then be forced to have probability 1, since* $\sum_{i \in I} P(E_i) = P(\cup_{i \in I} E_i) = 1$.] *Equality occurs at the upper extreme if and only if the events in* $\mathcal{E}$ *are equally likely.* [*In this case, each event in* $\mathcal{E}$ *would have probability* $1/|I|$.]

Proof: It is straightforward to see that the given conditions for equality at the two extremes are sufficient.

Since $0 \le P(E_i) \le 1$ for each $i \in I$, $-P(E_i)\log P(E_i) \ge 0$ with equality if and only if either $P(E_i) = 0$ (by convention) or $P(E_i) = 1$. Thus $H(\mathcal{E}) \ge 0$, and equality forces $P(E_i) = 0$ or 1 for each i. Since the E_i are pairwise mutually exclusive, and $\sum_i P(E_i) = 1$, $H(\mathcal{E}) = 0$ implies that exactly one of the E_i has probability 1 and the rest have probability zero.

Let $c = \log e$. We have

$$\begin{aligned}
H(\mathcal{E}) - \log|I| &= \sum_{i\in I} P(E_i)\log\frac{1}{P(E_i)} - \sum_{i\in I} P(E_i)\log|I| \\
&= c\sum_{i\in I} P(E_i)\ln(P(E_i)|I|)^{-1} \\
&\le c\sum_{i\in I} P(E_i)[(P(E_i)|I|)^{-1} - 1] \quad \text{(by Lemma 2.1.1)} \\
&= c\left[\sum_{i\in I} |I|^{-1} - \sum_{i\in I} P(E_i)\right] = c[1-1] = 0,
\end{aligned}$$

with equality if and only if $P(E_i)|I| = 1$ for each $i \in I$. □

The following theorem gives a useful connection between conditional entropy and the set-wise relation between two systems. Notice that if $\mathcal{E}$ is an amalgamation of $\mathcal{F}$, then whenever you know which event in $\mathcal{F}$ occurred, you also know which event in $\mathcal{E}$ occurred; i.e., there is no uncertainty regarding $\mathcal{E}$.

2.3.5 Theorem *Suppose that $\mathcal{E} = [E_i; i \in I]$ and $\mathcal{F} = [F_j; j \in J]$ are systems of events in some finite probability space. Then $H(\mathcal{E}|\mathcal{F}) = 0$ if and only if $\mathcal{E}$ is an amalgamation of $\mathcal{F}$.*

Proof: $H(\mathcal{E}|\mathcal{F}) = \sum_{i\in I}\sum_{j\in J} P(E_i \cap F_j)\log\frac{P(F_j)}{P(E_i\cap F_j)} = 0 \Leftrightarrow$ for each $i \in I, j \in J$, $P(E_i \cap F_j)\log\frac{P(F_j)}{P(E_i\cap F_j)} = 0$, since the terms of the sum above are all non-negative.

If $P(F_j) = 0$ then $P(E_i \cap F_j) = P(F_j)$ for any choice of $i \in I$. Since $P(F_j) = \sum_{i\in I} P(E_i \cap F_j)$ by Corollary 2.2.8, if $P(F_j) > 0$ then $P(E_i \cap F_j) > 0$ for some $i \in I$, and then $P(E_i \cap F_j)\log\frac{P(F_j)}{P(E_i\cap F_j)} = 0$ implies $P(E_i \cap F_j) = P(F_j)$. Thus $H(\mathcal{E}|\mathcal{F}) = 0$ implies that $\mathcal{E}$ is an amalgamation of $\mathcal{F}$, and the converse is straightforward to see. □

Exercises 2.3

1. Treating the sets of outcomes as systems of events, write out the entropies of each of $\mathcal{S}_1, \dots, \mathcal{S}_4$ in Exercise 2.2.1.

2. In the experiment of n independent Bernoulli trials with probability p of success on each trial, let E_k = "exactly k successes," and $\mathcal{E} = [E_0, \dots, E_n]$. Let $\mathcal{S} = \{S, F\}^n$, and treat $\mathcal{S}$ as a system of events (i.e., each element of $\mathcal{S}$,

regarded as an outcome of the experiment, is also to be thought of as an event). Write out both $H(\mathcal{E})$ and $H(\mathcal{S})$.

3. For a system $\mathcal{E}$ of events, show that $I(\mathcal{E},\mathcal{E}) = H(\mathcal{E})$.
4. (a) Show that, if $x_1,\ldots,x_n \ge 0$, then

$$\left(\sum_{i=1}^{n} x_i\right)\log\left(\sum_{i=1}^{n} x_i\right) \ge \sum_{i=1}^{n} x_i \log x_i .$$

   (b) Show that if $\widehat{\mathcal{E}}$ is obtained from $\mathcal{E}$ by amalgamation, then $H(\widehat{\mathcal{E}}) \le H(\mathcal{E})$.
5. Suppose that $\mathcal{E} = [E_i; i \in I]$ and $\mathcal{F} = [F_j; j \in J]$ are systems of events in some finite probability space. Under what conditions on $\mathcal{E}$ and $\mathcal{F}$ will it be the case that $H(\mathcal{E} \mid \mathcal{F}) = H(\mathcal{E})$? [See Theorem 2.4.4, next section.]
6. Suppose that $\mathcal{E}$ and $\mathcal{F}$ are systems of events in probability spaces associated with two (different) experiments. Suppose that the two experiments are performed *independently*, and the set of outcomes of the compound experiment is identified with $\mathcal{S}_1 \times \mathcal{S}_2$, where $\mathcal{S}_1$ and $\mathcal{S}_2$ are the sets of outcomes for the two experiments separately. Let

$$\mathcal{E}\cdot\mathcal{F} = [E \times F;\ E \in \mathcal{E}, F \in \mathcal{F}].$$

   Verify that $\mathcal{E}\cdot\mathcal{F}$ is a system of events in the space of the compound experiment.

   Show that $H(\mathcal{E}\cdot\mathcal{F}) = H(\mathcal{E}) + H(\mathcal{F})$. Will this result hold (necessarily) if the two experiments are not independent?

2.4 Information and entropy

Throughout this section, $\mathcal{E}$ and $\mathcal{F}$ will be systems of events in some finite probability space.

2.4.1 Theorem $I(\mathcal{E},\mathcal{F}) = H(\mathcal{E}) + H(\mathcal{F}) - H(\mathcal{E}\wedge\mathcal{F})$.

Proof:

$$\begin{aligned} I(\mathcal{E},\mathcal{F}) &= \sum_i \sum_j P(E_i \cap F_j)\log\frac{P(E_i\cap F_j)}{P(E_i)P(F_j)} \\ &= \sum_i \sum_j P(E_i \cap F_j)\log P(E_i \cap F_j) \\ &\qquad - \sum_i \sum_j P(E_i \cap F_j)\log P(E_i) - \sum_i \sum_j P(E_i \cap F_j)\log P(F_j) \end{aligned}$$

$$= -H(\mathcal{E}\wedge\mathcal{F}) - \sum_i P(E_i)\log P(E_i)$$

$$- \sum_j P(F_j)\log P(F_j) \quad \text{[using Corollary 2.2.8]}$$

$$= -H(\mathcal{E}\wedge\mathcal{F}) + H(\mathcal{E}) + H(\mathcal{F}). \qquad \square$$

2.4.2 Corollary $I(\mathcal{E},\mathcal{F}) \le H(\mathcal{E}) + H(\mathcal{F})$.

2.4.3 Corollary $H(\mathcal{E}\wedge\mathcal{F}) \le H(\mathcal{E}) + H(\mathcal{F})$, *with equality if and only if* $\mathcal{E}$ *and* $\mathcal{F}$ *are statistically independent.*

2.4.4 Theorem $H(\mathcal{E} \mid \mathcal{F}) = H(\mathcal{E}\wedge\mathcal{F}) - H(\mathcal{F}) = H(\mathcal{E}) - I(\mathcal{E},\mathcal{F})$.

2.4.5 Corollary $H(\mathcal{E} \mid \mathcal{F}) \le H(\mathcal{E})$, *with equality if and only if* $\mathcal{E}$ *and* $\mathcal{F}$ *are statistically independent.*

2.4.6 Corollary $I(\mathcal{E},\mathcal{F}) \le \min(H(\mathcal{E}), H(\mathcal{F}))$.

Proof: It will suffice to see that $I(\mathcal{E},\mathcal{F}) \le H(\mathcal{E})$. This follows from Theorem 2.4.4 and the observation that $H(\mathcal{E} \mid \mathcal{F}) \ge 0$. □

Notice that Corollary 2.4.6 is much stronger than Corollary 2.4.2.

Exercises 2.4

1–4. Prove 2.4.2, 2.4.3, 2.4.4, and 2.4.5, above.

5. From 2.4.1 and 2.3.4 deduce necessary and sufficient conditions on $\mathcal{E}$ and $\mathcal{F}$ for $I(\mathcal{E},\mathcal{F}) = H(\mathcal{E}) + H(\mathcal{F})$.

6. Express $H(\mathcal{E}\wedge\mathcal{E})$ and $H(\mathcal{E} \mid \mathcal{E})$ as simply as possible.

7. Three urns, A, B, and C, contain colored balls, as follows:

   A contains three red and five green balls,
   B contains one red and two green balls, and
   C contains seven red and six green balls.

   An urn is chosen, at random, and then a ball is drawn from that urn. Let the urn names also stand for the event that that urn was chosen, and let R = "a red ball was chosen," and G = "a green ball was chosen." Let $\mathcal{E} = \{A, B, C\}$ and $\mathcal{F} = \{R, G\}$. Write out $I(\mathcal{E},\mathcal{F})$, $H(\mathcal{E})$, $H(\mathcal{F})$, $H(\mathcal{E}\wedge\mathcal{F})$, $H(\mathcal{E} \mid \mathcal{F})$, and $H(\mathcal{F} \mid \mathcal{E})$. If, at any stage, you can express whatever you are trying to express in terms of items already written out, do so.

8. Regarding Corollary 2.4.6: under what conditions on $\mathcal{E}$ and $\mathcal{F}$ is it in the case that $I(\mathcal{E},\mathcal{F}) = H(\mathcal{E})$?

9. With the urns of problem 7 above, we play a new game. First draw a ball from urn A; if it is red, draw a ball from urn B; if the ball from urn A is green, draw a ball from urn C. Let $\mathcal{E}$ and $\mathcal{F}$ be the first and second

stage systems of events for this two-stage experiment; i.e., $\mathcal{E} = [R, G]$ and $\mathcal{F} = [\widetilde{R}, \widetilde{G}]$, where, for instance, $R =$ "the first ball drawn was red" and $\widetilde{R} =$ "the second ball drawn was red."

(a) Write out $I(\mathcal{E}, \mathcal{F})$, $H(\mathcal{E})$, $H(\mathcal{F})$, $H(\mathcal{E} \wedge \mathcal{F})$, $H(\mathcal{E} \mid \mathcal{F})$, and $H(\mathcal{F} \mid \mathcal{E})$ in this new situation.

(b) Suppose now that you are allowed to transfer balls between urns B and C. How would you rearrange the balls in those urns to maximize $I(\mathcal{E}, \mathcal{F})$? What is that maximum value?

(c) How would you rearrange the balls in urns B and C to minimize $I(\mathcal{E}, \mathcal{F})$? What is that minimum value?

(d) Answer the same questions in (b) and (c) with $I(\mathcal{E}, \mathcal{F})$ replaced by $H(\mathcal{E} \mid \mathcal{F})$.

(e) Under which of the rearrangements you produced in (b), (c), and (d) is $\mathcal{E}$ an amalgamation of $\mathcal{F}$? Under which is $\mathcal{F}$ an amalgamation of $\mathcal{E}$? Under which are $\mathcal{E}$ and $\mathcal{F}$ statistically independent?