Chapter 2

A little bit of information theory

2.1 Hartley's postulate

Information becomes a scientific notion only if made measurable. According to Webster's Dictionary

> ***bit***: binary digit, but also *unit* of information, or *elementary carrier* of information, or even a small quantity of food; esp: a small delicacy.

In order to avoid confusion, we call the classical physics elementary *carrier of information* not a bit, but a *cebit*.

A **cebit** is a physical thing which exists in only two configutations □ and ■. An example would be a coin on the table, the two possible configurations being "Head up" and "Tail up". We may assign a symbolic value to each configuration

$$b(\square) = 0\,, \qquad b(\blacksquare) = 1\,, \tag{2.1}$$

which defines a *dichotomic variable* b – the *bit* – the possible values of which are the **binary digits** `0` and `1`.[1] Here the distinction between the cebit and the bit seems a little artificial, but it turns out to be useful in quantum information theory.

In flipping a coin, you don't know whether the coin will land with head up or tail up on the table. Before the experiment you are just ignorant about its result, but once it is finished, and you know the result, your ignorance is gone. The information theorist view of ignorance is that of uncertainty about the outcome of a statistical experiment which has a finite number of possible outcomes. For the particular case of an experiment with two possible outcomes which are equally likely to occur (like in flipping a fair coin), Hartley postulates that

The outcome of an unbiased alternative provides one bit of information . (2.2)

which defines the third meaning of "bit" as a **unit of information**.

A statistical experiment where all outcomes are equally likely to occur is called a **Laplace experiment**. A particular outcome of such experiment supplies the same amount of information as any other outcome, namely

$$I = \mathrm{ld}(M)\ \mathrm{bit} \tag{2.3}$$

where M is the total number of possible outcomes and ld the **logarithm dualis**, that is the logarithm to base 2. For flipping a fair coin, for example, $M = 2$, and thus $I = 1\mathrm{bit}$ in accordance with (2.2). The flipping of N fair coins has $M = 2^N$ possible outcomes. The information supplied by any particular outcome is $I = N\mathrm{bit}$ which reflects the fact that you need to resolve N alternatives in order to specify a string of N binary digits.

[1] Please recall: digits are not numbers. Without additional conventions, you just can't add/subtract/multiply/divide digits.

2.2 Shannon Entropy

If you receive 100 cebits, and each of the cebits is equally likely □ or ■, the information provided would be 100 bits. If, however, you know their values prior to reading, the information provided would be 0 bits (the cebits would not provide any information). The Shannon Entropy quantifies the information between these two extremes.

The key concept is a **discrete source**, that is a process which generates a stream of **symbols**, $\dots x_{-1}, x_0, x_1, x_2, \dots$ where each x_t an element of a pre-defined set,

$$\mathcal{A} = \{a_1, a_2, \dots, a_A\}\,, \tag{2.4}$$

called an **alphabet** of A symbols. A **message** is any finite string of symbols

$$\mathbf{x} = x_{t_1} \cdots x_{t_N}\,. \tag{2.5}$$

The set of all strings of a fixed length N, denoted $\mathcal{A}^N$, is the N-fold cartesian product $\mathcal{A}^N = \mathcal{A} \times \mathcal{A} \times \cdots \times \mathcal{A}$.[2] The collection of all finite strings over an alphabet $\mathcal{A}$ is denoted $\mathcal{A}^*$.

In the Shannon-Model the x_t are identically distributed, independent random variables, (short: i.d.i. variables), that is

$$\text{Prob}(x_t = a_i) =: p_i \equiv p(a_i)\,, \qquad \sum_{i=1}^{A} p_i = 1\,. \tag{2.6}$$

A source of this kind is called a **stationary memoryless source**, abbreviated $X = (\mathcal{A}, \mathbf{p})$, where $\mathcal{A}$ ist the source's alphabet, and $\mathbf{p}$ the array of the values of the probability mass function on $\mathcal{A}$, also called **probability distribution**, $\mathbf{p} = (p_1, \dots, p_A)$.

[2] Alternatively: Alphabet is a set of letters (or characters) and a sequence of letters is called a word.

Indeed, according to (2.6) the probability distribution of x_t is independent of time t (the source is stationary) and also independent of the history or future of outcomes (the source is memoryless).

A message is just a finite string of the values of identically distributed, independent random variables. For sufficiently large N, the law of large numbers promises that, with probability arbitrarily close to one, the symbol a_i occurs $N_i \approx Np_i$ times. A string for which this is the case is called **typical string**. The number of typical strings is $M_{\text{typ}} \approx \frac{N!}{\prod_i (Np_i)!}$. Within the class of typical strings, the probability to encounter any particular string is uniform $P(\mathbf{x}|\mathbf{x} \text{ is typical}) \approx 1/M_{\text{typ}}$.[3] Following Hartley, Eq. (2.3), the information of a typical string, measured in bits, is given by $I = \text{ld} M_{\text{typ}}$. Since N is large we may use Stirling's approximation with the result

$$I = NH(X)\,, \tag{2.7}$$

where $H(X)$ is the **Shannon entropy**

$$H(X) = -\sum_{i=1}^{A} p_i \text{ld}(p_i)\,, \tag{2.8}$$

measured in bits. It is sometimes written $H[\mathbf{p}]$, which reflects the fact that, for given alphabet $\mathcal{A}$, it is just a functional on the space of probability mass functions.

The Shannon entropy is characteristic for a given source $X = (\mathcal{A}, \mathbf{p})$. It gives the **average** amount of information, measured in bits, carried by each symbol emitted by the source. If the transmission from source to sink is free of noise (ideal channel), it also gives the average amount of uncertainty removed per symbol received.

[3] More precisely: We define $\mathbf{x}$ is δ-typical $\Leftrightarrow H(X) - \delta < -\frac{1}{N}\text{ld}p(\mathbf{x}) < H(X) + \delta$. Furthermore $\forall \delta > 0, N \geq 1, \exists \varepsilon > 0 : \text{Prob}(\mathbf{x} \text{ is } \delta\text{-typical}|\mathbf{x}) > 1 - \varepsilon$ with $\lim_{N\to\infty} \varepsilon \to 0$, i.e. the probability that *any* string drawn from X^N is in fact typical becomes arbitrarily close to 1 in the limit of large N.

For the particular case of a **binary source** with alphabet $\mathcal{A} = \{\mathtt{H}, \mathtt{T}\}$, the Shannon entropy

$$H(p) = -p\,\mathrm{ld}(p) - (1-p)\,\mathrm{ld}(1-p)\,, \tag{2.9}$$

where p is the probability that – at a given instance – the source emits H, and $q := 1-p$ the probability that it emits T. The function (2.9) is sometimes called the **binary entropy function**. It is depicted in Fig. (2.1). Clearly, for small p – that is, when it is a-priori almost certain that the symbol reads T, not much information is revealed on average. The same holds in the opposite case $p \approx 1$, when it is a-priori certain, that the symbol reads H. The maximum is attained for equal probabilties $p = q = \frac{1}{2}$ in which case the cebit carries one bit of information in accordance with (2.2).

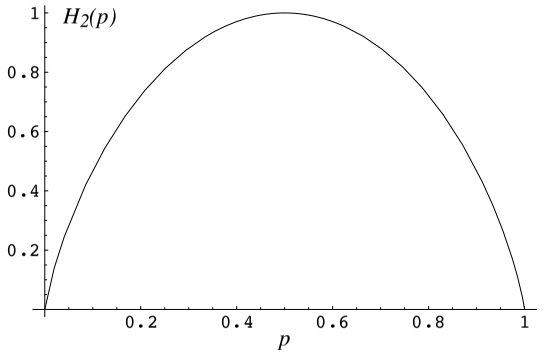

Fig. 2.1 *The binary entropy function (2.9).*

Theorem The Shannon entropy H for probabilities on A symbols is bounded

$$0 \leq H(X) \leq H_0(X) \equiv \mathrm{ld}A \tag{2.10}$$

and assumes its maximum $H_0(X)$ – called the **raw bit content** of X – for $\mathbf{p}$ uniform, $p_i = 1/A$.

The proof is simple: Rewrite $H = -\sum_{i=1}^{A-1} p_i \mathrm{ld}p_i - p_A \mathrm{ld}p_A$. Observing $p_A = 1 - \sum_{i=1}^{A-1} p_i$ solve $\partial H/\partial p_i = 0$ for $i = 1, \ldots, A-1$. Find $p_i = p_A = 1/A$ with corresponding $H = \mathrm{ld}A$. Compute the second derivative and confirm that H is indeed maximal for p uniform. *End-of-Proof*

The difference $H_0(X) - H(X)$ measures the **compressibility** of a source text. Adopting a block code which assigns integers to the strings of N symbols emitted by a memoryless source with Shannon entropy $H(X)$, you would need to convey only $NH(X)$ binary digits – instead of $NH_0(X)$ bits – in order to specify any string of

N symbols asymptotically, i.e. in the limit $N \to \infty$. The block code misses the untypical strings, but as these occur, for N large, with small probability, the chances of a coding error can be made arbitrarily small for sufficiently large N.

The Shannon entropy is easily identified with the expectation value of a stochastic variable

$$h(x) = \mathrm{ld}\frac{1}{p(x)} \tag{2.11}$$

called the **(Hartley) information content** of the outcome x. For Laplace experiments, the Hartley information coincides with the raw bit content in accordance with (2.3). But if the probability distribution is not uniform, symbols with low probability carry more bits than the raw bit content, while symbols with high probability carry less bits than the raw bit content. A rare event comes with "great surprise" and thus is very informative. On the other hand, a frequent event is "no surprise" and thus is only little informative. This imbalance of the Hartley information finds nice application in the construction of compact codes (for codes and such see the supplement 3.4).

Example: Consider for example a source X over an alphabet of four symbols $\mathcal{A} = \{a, b, c, b\}$ which are produced with probabilites

$$p(a) = \frac{1}{2}\,, \quad p(b) = \frac{1}{4}\,, \quad p(c) = p(d) = \frac{1}{8}\,. \tag{2.12}$$

The corresponding Shannon is $H = 1.75$bit. We intend to encode the symbols into strings of the binary alphabet $\mathcal{B} = \{\mathtt{0}, \mathtt{1}\}$. Comparing the "naive" code

$$f_1: \qquad a \to 00\,, \quad b \to 01\,, \quad c \to 10\,, \quad d \to 11\,, \tag{2.13}$$

and the "smart" code

$$f_2: \qquad a \to 1\,, \quad b \to 01\,, \quad c \to 001\,, \quad d \to 000\,, \tag{2.14}$$

which has the property, that the most probable symbol is assigned the shortes bit string, while the less probable symbols are assigned longer strings. Evidently, the length of the binary string wich encodes a given symbol $x \in \mathcal{A}$ coincides with the Hartley information carried by that symbol. The average word length, $\langle f \rangle := \sum_{x\in\mathcal{A}} p(x)|f(x))|$, for the smart code is shorter than for the naive code, and is just the Shannon $\langle f_2 \rangle = 1.75\text{bit} < \langle f_1 \rangle = 2\text{bit}$. The "smart" is uniquely decipherable and instantaneous, that is concatenations of code words can be decoded 'on line' without looking into the future.

Mathematically, the Shannon entropy is a functional on the set of probability distributions: it measures the degree of "spreadiness" of a distribution. And spreadiness increases with increasing uncertainty. If you are uncertain which probability distribution $\mathbf{p}$ or $\mathbf{q}$ is characteristic for your source, the corresponding Shannon is larger than the average. This is the meaning of the

Theorem The Shannon entropy H is **concave** on the set $\mathcal{P} = \{\mathbf{p} : \mathbf{p} \text{ is a probability distribution}\}$, that is

$$H(\lambda\mathbf{p} + (1-\lambda)\mathbf{q}) \geq \lambda H(\mathbf{p}) + (1-\lambda)H(\mathbf{q})\,, \quad 0 \leq \lambda \leq 1\,. \tag{2.15}$$

The proof is simple. Let $r_i = \lambda p_i + (1-\lambda)q_i$. Use Gibb's inequality (see below), and rewrite the right hand side, $0 \geq -\lambda H(\mathbf{p}\|\mathbf{q}) - (1-\lambda)H(\mathbf{q}\|\mathbf{r}) = \ldots = \lambda H(\mathbf{p}) + (1-\lambda)H(\mathbf{q}) - \sum r\text{ld}r$. *End-of-proof* Concavity, and her sister convexity, play an important role in information theory and statistical physics – see the Complement A.4.

2.3 Relative entropy

The relative entropy is defined

$$H(\mathbf{p}\|\mathbf{q}) := \sum_{i=1}^{N} p_i \mathrm{ld}\frac{p_i}{q_i}\,. \tag{2.16}$$

It obeys the **Gibbs inequality**

$$H(\mathbf{p}\|\mathbf{q}) \geq 0\,, \tag{2.17}$$

with equality only if $\mathbf{p} = \mathbf{q}$.

The proof is easy. We have

$$\begin{aligned} H(\mathbf{p}\|\mathbf{q}) &= -\sum_j p_j \mathrm{ld}\frac{q_j}{p_j} && (2.18)\\ &= -\sum_j \left(p_j \mathrm{ld}\frac{q_j}{p_j} - \frac{q_j - p_j}{\log 2} \right) && (2.19)\\ &\geq 0\,, && (2.20) \end{aligned}$$

where the first equation is based on $\mathrm{ld}x = -\mathrm{ld}(1/x)$, the second equation is based on $\sum_j p_j = \sum_j q_j = 1$, and the final inequality is based on the concavity of log-functions (bounded from above by any of its tangent)

$$\mathrm{ld}x = \frac{\log x}{\log 2} \leq \frac{x-1}{\log 2} \tag{2.21}$$

with equality if and only if $x = 1$. *End-of-proof*

The relative entropy is also called the **Kullback-Leibler** divergence between the probability distributions $\mathbf{p}$ and $\mathbf{q}$. It measures how many bits per symbol are wasted by using a code which is optimized for an assumed probability distribution $\mathbf{q}$ if the "true" distribution is not $\mathbf{q}$ but rather $\mathbf{p}$.[4]

2.4 Complement: Convex analysis – the basics

Definition A function $f(x)$ is *convex* over $[a, b] \subset \mathbb{R}$ if, for all $x_1, x_2 \in [a, b]$ and $0 \leq \lambda \leq 1$,

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda) f(x_2)\,. \tag{2.22}$$

A function f is *strictly convex* if, for all $x_1, x_2 \in [a, b]$, equality holds only for $\lambda = 0$ and $\lambda = 1$.

Similar definitions apply also to *concave* and *strictly concave* functions. A mnemonic for convexity: if you say "convex", you are likely to smile – the corners of you mouth go up.

Theorem If f is a convex function of a random variable x, then the expectation value of f with respect to any given probability distribution $p(x)$ obeys the *Jensen's inequality*

$$\langle f \rangle \geq f(\langle x \rangle)\,. \tag{2.23}$$

[4]Have a look at the source coding theorem. Take a variable length code C_q where codeword i has length $\approx 1/\mathrm{ld}(q_i)$. For a source with probability distribution $\mathbf{p}$, the average length of this code is $L(C_q, X) \approx \sum p_i \mathrm{ld}(1/q_i)$. This is larger than the average length of the optimal code C_p, which is $L(C_p, X) \approx H(X) = \sum p_i \mathrm{ld}(1/p_i)$. In fact $H(\mathbf{p}\|\mathbf{q}) = L(C_q, X) - L(C_p, X) \geq 0$, see Eq. (2.17).

Here comes the proof: The center of gravity of a freely hanging massive chord lies always above the chord. *End-of-proof*

Convex set is a set $\mathcal{P}$ such that for every pair $\mathbf{p}_1, \mathbf{p}_2 \in \mathcal{P}$, the convex combination

$$\mathbf{p}_{12} = \lambda \mathbf{p}_1 + (1-\lambda)\mathbf{p}_2\,, \quad 0 \leq \lambda \leq 1 \tag{2.24}$$

belongs to the set, $\mathbf{p}_{12} \in \mathcal{P}$.

Important Theorem I The set of probability distributions is a convex set.

Convex Hull of a set of points $\mathcal{S}$ is the intersection of all convex sets containing $\mathcal{S}$.

Carathéodory's Fundamental Theorem states that each point in the convex hull of a set $\mathcal{S} \subset \mathbb{R}^n$ is in the convex combination of $H_n = n+1$ or fewer points of $\mathcal{S}$. H_n is called the Helly Number of Euclidean n-space $\mathbb{R}^n$.

Helly's Theorem states that if F is a family of more than n bounded closed convex sets in Euclidean n-space $\mathbb{R}^n$, and if every H_n members of F have at least one point in common, then all the members of F have at least one point in common.

Important Theorem II Every probability distribution has a unique decomposition into extremal points, called pure states.

Important Remark The "Imporatnt Theorem II" does not hold for quantum mechanical density operators.

