# Information Theory

University of Amsterdam, Master of Logic, Fall 2016

Yfke Dulek and Christian Schaffner

The source is available on github and we are grateful for any feedback including typos and suggestions for improvements! Feel free to send us pull requests.

Parts of the text is based on, or taken verbatim, from ‘The Mathematical Theory of Information, and Applications’ (version 2.0) by Ronald Cramer and Serge Fehr [CF]. The lay-out is based on the Legrand Orange Book (version 2.1.1), licenced under the Creative Commons License BY-NC-SA 3.0.

# Contents

# Week 1: Probability and Entropy

## 1.1 Preliminaries: Probability Theory

For this course, we will only be concerned with discrete probabilities. This section formalizes some notions you should already be familiar with: probability spaces, events and probability distributions.

**Definition 1.1.1 — Probability space.** A (discrete) probability space $(\Omega, \mathcal{F}, P)$ consists of a discrete, non-empty *sample space* $\Omega$, an *event space* $\mathcal{F} \subseteq \mathcal{P}(\Omega)$ and a *probability measure* $P$ which is a function $P : \Omega \to \mathbb{R}_{\geq 0}$ that satisfies

$$\sum_{\omega \in \Omega} P(\omega) = 1.$$

The event space $\mathcal{F}$ is required to be non-empty and closed under intersection, union and complements. For convenience, we will most often assume that $\mathcal{F}$ equals the powerset $\mathcal{P}(\Omega)$ of $\Omega$, i.e. it contains all possible subsets of events, and therefore fulfils the required properties.

**Definition 1.1.2 — Event.** An event $\mathcal{A}$ is an element of the event space $\mathcal{F} \subseteq \mathcal{P}(\Omega)$, i.e. a subset $\mathcal{A}$ of the sample space $\Omega$. Its probability is defined as

$$P[\mathcal{A}] := \sum_{\omega \in \mathcal{A}} P(\omega),$$

where by convention $P[\emptyset] = 0$.

As a notational convention, we write $P[\mathcal{A}, \mathcal{B}]$ for $P[\mathcal{A} \cap \mathcal{B}]$, and $P[\overline{\mathcal{A}}]$ for $P[\Omega \backslash \mathcal{A}]$.

**Exercise 1.1** Prove the following identities (for arbitrary events $\mathcal{A}, \mathcal{B} \subseteq \Omega$):

$$P[\overline{\mathcal{A}}] = 1 - P[\mathcal{A}] \tag{1.1}$$

$$P[\mathcal{A} \cup \mathcal{B}] = P[\mathcal{A}] + P[\mathcal{B}] - P[\mathcal{A}, \mathcal{B}] \tag{1.2}$$

$$P[\mathcal{A}] = P[\mathcal{A}, \mathcal{B}] + P[\mathcal{A}, \overline{\mathcal{B}}]. \tag{1.3}$$

■

It is often useful to consider the probability of an event *given* that some other event happened:

**Definition 1.1.3 — Conditional probability.** For events $\mathcal{A}$ and $\mathcal{B}$ with $P[\mathcal{A}] > 0$, the conditional probability of $\mathcal{B}$ given $\mathcal{A}$ is defined as

$$P[\mathcal{B}|\mathcal{A}] := \frac{P[\mathcal{A},\mathcal{B}]}{P[\mathcal{A}]}\,.$$

**Example 1.1 — Fair die.** We throw a six-sided fair die once, and consider the number that comes up. The sample space for this experiment is $\Omega = \{1,2,3,4,5,6\}$, with event space $\mathcal{F} = \mathcal{P}(\Omega)$ and probability measure $P[i] = \frac{1}{|\Omega|} = \frac{1}{6}$ for all $i \in \Omega$ (this is a **uniform** probability measure). Consider the events $\mathcal{A} = \{2,4,6\}$ and $\mathcal{B} = \{3,6\}$. Using the formulas in Definitions 1.1.2 and 1.1.3, we can compute the following probabilities:

$$\begin{aligned}
P[\mathcal{A}] &= \frac{1}{2} && \text{(the outcome is even)}\\
P[\mathcal{B}] &= \frac{1}{3} && \text{(the outcome is a multiple of 3)}\\
P[\mathcal{A},\mathcal{B}] &= P[\{6\}] = \frac{1}{6} && \text{(the roll is even \textit{and} a multiple of 3)}\\
P[\mathcal{A}|\mathcal{B}] &= \frac{1/6}{1/3} = \frac{1}{2} && \text{(the roll is even, \textit{given} that it is a multiple of 3)}\\
P[\mathcal{B}|\mathcal{A}] &= \frac{1/6}{1/2} = \frac{1}{3} && \text{(the roll is a multiple of 3, \textit{given} that it is even)}
\end{aligned}$$

This example shows that in general, $P[\mathcal{A}|\mathcal{B}]$ is *not equal* to $P[\mathcal{B}|\mathcal{A}]$.

**Definition 1.1.4 — Discrete Random Variable (RV).** Let $(\Omega,\mathcal{F},P)$ be a discrete probability space. A random variable $X$ is a function $X : \Omega \to \mathcal{X}$ where $\mathcal{X}$ is a set, and we may assume it to be discrete.

A *real* random variable is one whose image is contained in $\mathbb{R}$. A (The *image* and the *range* of a random variable $X$ are given by the image and the range of $X$ in the function-theoretic sense.) The image of a *binary* random variable is a set $\{x_0,x_1\}$ with only two elements.

**Definition 1.1.5 — Probability distribution.** Let $X$ be a random variable. The probability distribution of $X$ is the function $P_X : \mathcal{X} \to [0,1]$ defined as

$$P_X(x) := P[X = x],$$

where $X = x$ denotes the event $\{\omega \in \Omega \,|\, X(\omega) = x\}$.

Alternatively, one can write $P_X(x) = P(X^{-1}(x))$ to express that the probability of $x$ is precisely the $P$-measure of the pre-image of $x$ under the random variable $X$.

**Exercise 1.2** Verify that $(\mathcal{X},\mathcal{P}(\mathcal{X}),P_X)$ is itself a probability space. ■

We say that $P_X$ is a **uniform** distribution if the associated probability measure is uniform, i.e. $P_X(x) = \frac{1}{|\mathcal{X}|}$. The **support** of a random variable or a probability distribution is defined as $\mathrm{supp}(P_X) := \{x \in \mathcal{X} \mid P_X(x) > 0\}$, the points of the range which have strictly positive probability. We often slightly abuse notation and write $\mathrm{supp}(X)$ instead.

When given two or more random variables defined on the same probability space, we can consider the probability that each of the variables take on a certain value:

**Definition 1.1.6 — Joint probability distribution.** Let $X$ and $Y$ be two random variables defined on the same probability space, with respective ranges $\mathcal{X}$ and $\mathcal{Y}$. The pair $XY$ is a random variable with probability distribution $P_{XY} : \mathcal{X} \times \mathcal{Y} \to [0,1]$ given by

$$P_{XY}(x,y) := P[X = x, Y = y].$$

This definition naturally extends to three and more random variables. Unless otherwise stated, a collection of random variables is assumed to be defined on the same (implicit) probability space, so that their joint distribution is always well-defined.

If $P_{XY} = P_X \cdot P_Y$, in the sense that $P_{XY}(x,y) = P_X(x)P_Y(y)$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, then the random variables $X$ and $Y$ are said to be **independent**. If a set of variables $X_1, \ldots, X_n$ are all mutually independent and all have the same distribution (i.e. $P_{X_i} = P_{X_j}$ for all $i, j$), then they are **independent and identically distributed**, or **i.i.d.**

From a joint distribution, we can always find out the “original” (or **marginal**) distribution of one of the random variables (for example, $X$) by **marginalizing** out the variable that we want to discard (for example, $Y$):

$$P_X(x) = \sum_{y \in \mathcal{Y}} P_{XY}(x,y). \tag{1.4}$$

This marginalization process also works with more than two random variables.

Like events, probability distributions can also be conditioned on probabilistic events:

**Definition 1.1.7 — Conditional probability distribution.** If $\mathcal{A}$ is an event with $P[\mathcal{A}] > 0$, then the conditional probability distribution of $X$ given $\mathcal{A}$ is given by

$$P_{X|\mathcal{A}}(x) = \frac{P[X = x, \mathcal{A}]}{P[\mathcal{A}]}.$$

If $Y$ is another random variable and $P_Y(y) > 0$, then we write

$$P_{X|Y}(x|y) := P_{X|Y=y}(x) = \frac{P_{XY}(x,y)}{P_Y(y)}$$

for the conditional distribution of $X$, given $Y = y$.

Note that again, both $(\mathcal{X}, P_{X|\mathcal{A}})$ and $(\mathcal{X}, P_{X|Y=y})$ themselves form probability spaces. Note also that if $X$ and $Y$ are independent, then

$$P_{X|Y}(x|y) = \frac{P_{XY}(x,y)}{P_Y(y)} = \frac{P_X(x) \cdot P_Y(y)}{P_Y(y)} = P_X(x), \tag{1.5}$$

which aligns well with our intuition of independent variables: the distribution of $X$ remains unchanged when $Y$ is fixed to a specific value.

**Example 1.2 — Fair die (continued).** Consider again the throw of a six-sided fair die as in Example 1.1. Let the random variable $X$ describe the number of integer divisors for the outcome, that is

$$X(1) = 1 \quad X(2) = 2 \quad X(3) = 2 \quad X(4) = 3 \quad X(5) = 2 \quad X(6) = 3$$

$X$ is a real random variable, with range $\mathcal{X} = \{1,2,3\}$. The associated probability distribution is

$$P_X(1) = P[\{1\}] = \frac{1}{6}, \qquad P_X(2) = P[\{2,3,5\}] = \frac{1}{2}, \qquad P_X(3) = P[\{4,6\}] = \frac{1}{3}.$$

If we now condition on the event $\mathcal{A} = \{2,4,6\}$ (the outcome being even), we get that

$$P_{X|\mathcal{A}}(1) = 0, \qquad P_{X|\mathcal{A}}(2) = \frac{1}{3}, \qquad P_{X|\mathcal{A}}(3) = \frac{2}{3}.$$

If $X$ is a random variable and $f : \mathcal{X} \to \mathcal{Y}$ is a surjective function, then $f(X)$ is a random variable, defined by composing the map $f$ with the map $X$. Its image is $\mathcal{Y}$. Clearly,

$$P_{f(X)}(y) = \sum_{x \in \mathcal{X} : f(x) = y} P_X(x). \tag{1.6}$$

For example, $1/P_X(X)$ denotes the real random variable obtained from another random variable $X$ by composing with the map $1/P_X$ that assigns $1/P_X(x) \in \mathbb{R}$ to $x \in \mathcal{X}$.

**Definition 1.1.8 — Expectation.** The expectation of a *real* random variable $X$ is defined as

$$\mathbb{E}[X] := \sum_{x \in \mathcal{X}} P_X(x) \cdot x.$$

Note that if $X$ is not real, then we can still consider the expectation of some function $f : \mathcal{X} \to \mathbb{R}$, where

$$\mathbb{E}[f(X)] = \sum_{x \in \mathcal{X}} P_X(x) \cdot f(X). \tag{1.7}$$

**Definition 1.1.9 — Variance.** The variance of a *real* random variable $X$ is defined as

$$\mathrm{Var}[X] := \mathbb{E}[(X - \mathbb{E}[X])^2].$$

The variation is a measure for the deviation of the mean. Hoeffding's inequality (here stated for binary random variables) states that for a list of i.i.d. random variables, the average of the random variables is close to the expectation, except with very small probability. We state it here without proof.

**Theorem 1.1.1 — Hoeffding's inequality.** Let $X_1,...,X_n$ be independent and identically distributed binary random variables with $P_{X_i}(0) = 1 - \mu$ and $P_{X_i}(1) = \mu$, and thus $\mathbb{E}[X_i] = \mu$. Then, for any $\delta > 0$

$$P\left[\sum_i X_i > (\mu + \delta) \cdot n\right] \leq \exp(-2\delta^2 n).$$

## 1.2 Some Important Distributions

- The distribution of a biased coin with probability $P_X(1) = p$ to land heads, and a probability of $P_X(0) = 1 - p$ to land tails is called **Bernoulli($p$)** distribution. Its entropy is given by the binary entropy $h(p)$. The expected value is $\mathbb{E}[X] = p$ and the variance is $\mathrm{Var}[X] = p(1-p)$.
- When $n$ coins $X_1, X_2, \ldots, X_n$ are flipped independently and every $X_i$ is Bernoulli($p$) distributed, let $S = \sum_{i=1}^n X_i$ be their sum, i.e. the number of heads in $n$ throws of a biased coin. Then, $S$

has the **binomial($n,p$)** distribution:

$$P_S(k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad \text{where } k = 0,1,2,\dots,n. \tag{1.8}$$

From simple properties of the expected value and variance, one can show that $\mathbb{E}[S] = np$ and $\mathrm{Var}[S] = np(1-p)$.

- The **geometric($p$)** distribution of a random variable $Y$ is defined as the number of times one has to flip a Bernoulli($p$) coin before it lands heads:

$$P_Y(k) = (1-p)^{k-1} p \qquad \text{where } k = 1,2,3,\dots. \tag{1.9}$$

There is another variant of the geometric distribution used in the literature, where one excludes the final success event of landing heads in the counting:

$$P_Z(k) = (1-p)^k p \qquad \text{where } k = 0,1,2,3,\dots. \tag{1.10}$$

While the expected values are slightly different, namely $\mathbb{E}[Y] = \frac{1}{p}$ and $\mathbb{E}[Z] = \frac{1-p}{p}$, their variances are the same $\mathrm{Var}[Y] = \mathrm{Var}[Z] = \frac{1-p}{p^2}$.

## 1.3 Jensen's Inequality

In the following, let $\mathcal{D}$ be an interval in $\mathbb{R}$.

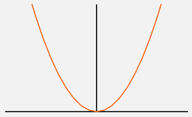
**Definition 1.3.1 — Convex and concave functions.** The function $f : \mathcal{D} \to \mathbb{R}$ is convex if for all $x_1, x_2 \in \mathcal{D}$ and all $\lambda \in [0,1] \subset \mathbb{R}$:

$$\lambda f(x_1) + (1-\lambda) f(x_2) \geq f(\lambda x_1 + (1-\lambda) x_2).$$

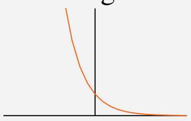
The function $f$ is *strictly* convex if equality only holds when $\lambda \in \{0,1\}$ or when $x_1 = x_2$. The function $f$ is (strictly) concave if the function $-f$ is (strictly) convex.

Intuitively, a function is convex if any straight line drawn between two points $f(x_1)$ and $f(x_2)$ lies above the graph of $f$ entirely. For a concave function, such a line must lie entirely beneath the graph.

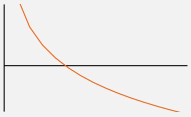
**Example 1.3** The following functions are convex (for $c \in \mathbb{R}$):



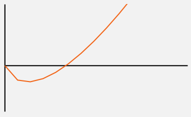
$x^2$



$e^{-x}$



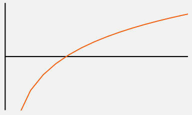
$\log(1/x)$



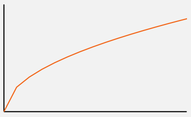
$x \cdot \log x$

$c$

The following functions are concave (for $c \in \mathbb{R}$):



$\log x$



$\sqrt{x}$

$c$

The following establishes a more formal method of proving the convexity of a function.

**Proposition 1.3.1** Let $f : \mathcal{D} \to \mathbb{R}$. If $\mathcal{D}$ is open, and for all $x \in \mathcal{D}$, the second order derivative $f''(x)$ exists and is non-negative (positive), then $f$ is convex (strictly convex).

We omit the proof, which can be found in, for example, [CF] (Lemma 1).

**Theorem 1.3.2 — Jensen's inequality.** Let $f : \mathcal{D} \to \mathbb{R}$ be a convex function, and let $n \in \mathbb{N}$. Then for any $p_1, \ldots, p_n \in \mathbb{R}_{\geq 0}$ such that $\sum_{i=1}^{n} p_i = 1$ and for any $x_1, \ldots, x_n \in \mathcal{D}$ it holds that

$$\sum_{i=1}^{n} p_i f(x_i) \geq f\left(\sum_{i=1}^{n} p_i x_i\right).$$

If $f$ is strictly convex and $p_1, \ldots, p_n > 0$, then equality holds iff $x_1 = \cdots = x_n$.

In particular, if $X$ is a real random variable whose image $\mathcal{X}$ is contained in $\mathcal{D}$, then

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X]),$$

and if $f$ is strictly convex, equality holds iff there is a $c \in \mathcal{X}$ such that $X = c$ with probability 1.

*Proof.* The proof is by induction. The case $n = 1$ is trivial, and the case $n = 2$ is identical to the very definition of convexity. Suppose that we have already proved the claim up to $n - 1 \geq 2$. Assume, without loss of generality, that $p_n < 1$. Then:

$$\begin{aligned}
\sum_{i=1}^{n} p_i f(x_i) &= p_n f(x_n) + \sum_{i=1}^{n-1} p_i f(x_i) \\
&= p_n f(x_n) + (1 - p_n) \sum_{i=1}^{n-1} \frac{p_i}{1 - p_n} f(x_i) \\
&\geq p_n f(x_n) + (1 - p_n) f\left(\sum_{i=1}^{n-1} \frac{p_i}{1 - p_n} x_i\right) && \text{(induction hypothesis)} \\
&\geq f\left(p_n x_n + (1 - p_n) \sum_{i=1}^{n-1} \frac{p_i}{1 - p_n} x_i\right) && \text{(definition of convexity)} \\
&= f\left(p_n x_n + \sum_{i=1}^{n-1} p_i x_i\right) \\
&= f\left(\sum_{i=1}^{n} p_i x_i\right). && (1.11)
\end{aligned}$$

That proves the claim. As for the strictness claim, if $x_1, \ldots, x_n$ are not all identical, then either $x_1, \ldots, x_{n-1}$ are not all identical and the first inequality is strict by induction hypothesis, or $x_1 = \cdots = x_{n-1} \neq x_n$ so that the second inequality is strict by the definition of convexity. ■

## 1.4 Shannon Entropy

In this section, we explore a measure for the amount of uncertainty of random variables. Consider some probabilistic event $\mathcal{A}$ that occurs with probability $P[\mathcal{A}]$ for some probability measure $P$. The **surprisal value** $\log \frac{1}{P[\mathcal{A}]}$ indicates how surprised we should be when the event $\mathcal{A}$ occurs: events with small probabilities yield high surprisal values, and vice versa. An event that occurs with certainty ($P_X(\mathcal{A}) = 1$) yields a surprisal value of 0. For a random variable $X$, we consider the *expected* surprisal value to be an indicator of how much uncertainty is contained in the variable, or how much information is gained by revealing the outcome. This expected surprisal value is more commonly known as the (Shannon) entropy[1] of a random variable:

**Definition 1.4.1 — Entropy.** Let $X$ be a random variable with image $\mathcal{X}$. The (Shannon) entropy $H(X)$ of $X$ is defined as

$$H(X) := \mathbb{E}\left[\log \frac{1}{P_X(X)}\right] = \sum_{x \in \mathcal{X}} P_X(x) \cdot \log \frac{1}{P_X(x)} = -\sum_{x \in \mathcal{X}} P_X(x) \cdot \log P_X(x),$$

with the convention that the log function represents the *binary* logarithm $\log_2$. As another convention, for $x \in \mathcal{X}$ with $P_X(x) = 0$, the corresponding argument in the summation is declared 0 (which is justified by taking a limit).

It is important to realize that the entropy of $X$ is a function (solely) of the *distribution* $P_X$ of $X$. However, it is customary to write $H(X)$ instead of the formally correct $H(P_X)$.

**Proposition 1.4.1 — Positivity.** Let $X$ be a random variable with image $\mathcal{X}$. Then

$$0 \leq H(X) \leq \log(|\mathcal{X}|).$$

Equality on the left-hand side holds iff there exists $x \in \mathcal{X}$ with $P_X(x) = 1$ (and thus $P_X(x') = 0$ for all $x' \neq x$). Equality on the right-hand side holds iff $P_X(x) = 1/|\mathcal{X}|$ for all $x \in \mathcal{X}$.

*Proof.* The function $f : \mathbb{R}_{>0} \to \mathbb{R}$ defined by $y \mapsto \log y$ is strictly concave on $\mathbb{R}_{>0}$. Thus, by Jensen's inequality:

$$H(X) = \sum_{x \in \mathcal{X}} P_X(x) \cdot \log \frac{1}{P_X(x)} \leq \log\left(\textstyle\sum_{x \in \mathcal{X}} 1\right) = \log(|\mathcal{X}|). \tag{1.12}$$

Furthermore, since we may restrict the sum to all $x$ with $P_X(x) > 0$, equality holds if and only if $\log(1/P_X(x)) = \log(1/P_X(x'))$, and thus $P_X(x) = P_X(x')$, for all $x, x' \in \mathcal{X}$.

Finally, for the characterization of the lower bound, it is obvious that $H(X) = 0$ if $P_X(x) = 1$ for some $x$, and, on the other hand, if $H(X) = 0$ then for any $x$ with $P_X(x) > 0$ it must be that $\log(1/P_X(x)) = 0$ and hence $P_X(x) = 1$. ■

For a binary random variable $X$ with image $\mathcal{X} = \{x_0, x_1\}$ and probabilities $P_X(x_0) = p$ and $P_X(x_1) = 1 - p$, we can write $H(X) = h(p)$, where $h$ denotes the binary entropy function:

**Definition 1.4.2 — Binary entropy function $h$.** The binary entropy function is defined for $0 < q < 1$ as

$$h(q) := q \log \frac{1}{q} + (1-q) \log \frac{1}{1-q},$$

and is defined as $h(q) = 0$ for $q = 0$ or $q = 1$. The graph of $h$ on the interval $[0, 1]$ is shown in Figure 1.1.

This binary entropy function is used, for example, to measure the entropy of a biased coin flip.

[1] Shannon once said: *My greatest concern was what to call it. I thought of calling it* information*, but the word was overly used, so I decided to call it* uncertainty*. When I discussed it with John von Neumann, he had a better idea. Von Neumann told me: "You should call it* entropy*, for two reasons. In the first place, your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, nobody knows what entropy really is, so in a debate you will always have the advantage."*

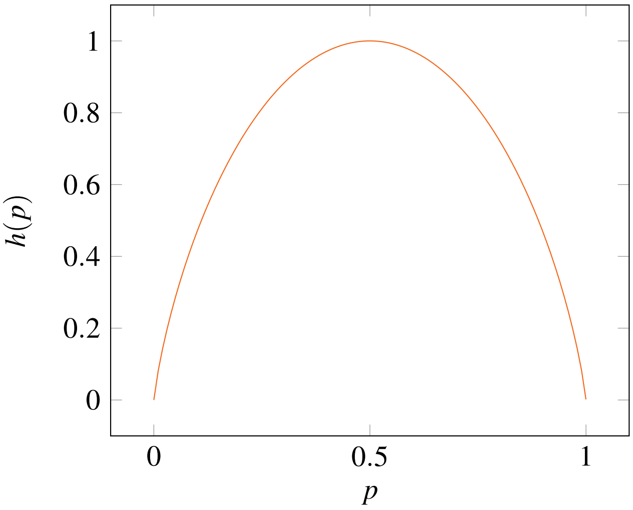


Figure 1.1: The binary entropy function $h$ as a function of the probability $p$.

**Example 1.4** Consider a random variable $X$ with $\mathcal{X} = \{a,b,c\}$ and $P_X(a) = \frac{1}{2}$, $P_X(b) = P_X(c) = \frac{1}{4}$. The entropy of $X$ is

$$H(X) = \frac{1}{2}\log 2 + \frac{1}{4}\log 4 + \frac{1}{4}\log 4 = \frac{1}{2} + \frac{1}{2} + \frac{1}{2} = \frac{3}{2}. \tag{1.13}$$

Another approach to computing the entropy of $X$ by coming up with an appropriate underlying probability space $(\Omega, P)$: we toss a fair coin twice, giving $\Omega = \{\mathsf{hh}, \mathsf{ht}, \mathsf{th}, \mathsf{tt}\}$ and $P(\omega) = \frac{1}{4}$ for all $\omega \in \Omega$. The we define the function $X : \Omega \to \mathcal{X}$ as

$$X(\mathsf{hh}) = X(\mathsf{ht}) = a, \qquad X(\mathsf{th}) = b, \qquad X(\mathsf{tt}) = c.$$

This yields the correct distribution $P_X$. The following computation now leads to the entropy of $X$:

$$H(X) = h\left(\frac{1}{2}\right) + \frac{1}{2}h(0) + \frac{1}{2}h\left(\frac{1}{2}\right) = \frac{3}{2}. \tag{1.14}$$

The first coin toss determines whether the outcome is $a$ (on heads $\mathsf{h}$) or something else (on tails $\mathsf{t}$). On heads, the second coin toss does not give any more information, whereas on tails, the second coin toss still decides between outcome $b$ and outcome $c$. In general, the entropy of a random variable with probabilities $p_1, \dots, p_n$ can be expressed as

$$\begin{aligned} H(p_1,\dots,p_k,p_{k+1},\dots,p_n) = h(p_1 + \cdots + p_k) + \\ (p_1 + \cdots + p_k) H\left(\frac{p_1}{p_1 + \cdots + p_k} + \cdots + \frac{p_k}{p_1 + \cdots + p_k}\right) + \\ (p_{k+1} + \cdots + p_n) H\left(\frac{p_{k+1}}{p_{k+1} + \cdots + p_n} + \cdots + \frac{p_n}{p_{k+1} + \cdots + p_n}\right). \end{aligned} \tag{1.15}$$

## 1.5 Conditional Entropy

Let $X$ be a random variable and $\mathcal{A}$ an event. Applying Definition 1.4.1 to the conditional probability distribution $P_{X|\mathcal{A}}$ allows us to naturally define the entropy of $X$ conditioned on the event $\mathcal{A}$, which leads to the following notion:

**Definition 1.5.1 — Conditional entropy.** Let $X$ and $Y$ be random variables, with respective images $\mathcal{X}$ and $\mathcal{Y}$. The conditional entropy $H(X|Y)$ of $X$ given $Y$ is defined as

$$H(X|Y) := \sum_{y\in\mathcal{Y}} P_Y(y)\cdot H(X|Y{=}y)\,,$$

with the convention that the corresponding argument in the summation is 0 for $y \in \mathcal{Y}$ with $P_Y(y) = 0$, and where

$$H(X|\mathcal{A}) := \sum_{x\in\mathcal{X}} P_{X|\mathcal{A}}(x)\cdot\log\frac{1}{P_{X|\mathcal{A}}(x)}\,.$$

Note that conditional entropy $H(X|Y)$ is not the entropy of a probability distribution but an expectation: the average uncertainty about $X$ when given $Y$. The following bound expresses that (on average!) additional information, i.e. knowing $Y$, can only *decrease* the uncertainty.

**Proposition 1.5.1** Let $X$ and $Y$ be random variables with respective images $\mathcal{X}$ and $\mathcal{Y}$. Then

$$0 \le H(X|Y) \le H(X)$$

Equality on the left-hand side holds iff $X$ is determined by $Y$, i.e., for all $y \in \mathcal{Y}$, there is an $x \in \mathcal{X}$ such that $P_{X|Y}(x|y) = 1$. Equality on the right-hand side holds iff $X$ and $Y$ are independent.

*Proof.* The lower bound follows trivially from the definition and from Proposition 1.4.1, and so does the characterization of when $H(X|Y) = 0$. For the upper bound, note that

$$H(X|Y) = \sum_y P_Y(y)\sum_x P_{X|Y}(x|y)\log\frac{1}{P_{X|Y}(x|y)} = \sum_{x,y} P_{XY}(x,y)\log\frac{P_Y(y)}{P_{XY}(x,y)} \tag{1.16}$$

and

$$H(X) = \sum_x P_X(x)\log\frac{1}{P_X(x)} = \sum_{x,y} P_{XY}(x,y)\log\frac{1}{P_X(x)} \tag{1.17}$$

where the last equality is derived by marginalization. Note that in both expressions, we may restrict the sum to those pairs $(x,y)$ with $P_{XY}(x,y) > 0$. Using Jensen's inequality, it follows that

$$\begin{aligned} H(X|Y) - H(X) &= \sum_{x,y} P_{XY}(x,y)\log\frac{P_X(x)P_Y(y)}{P_{XY}(x,y)} \\ &\le \log\Big(\sum_{x,y} P_X(x)P_Y(y)\Big) \le \log\Big(\big(\sum_x P_X(x)\big)\big(\sum_y P_Y(y)\big)\Big) = \log 1 = 0\,. \end{aligned} \tag{1.18}$$

Note that in the second inequality, we replaced the summation over all $(x,y)$ with $P_{XY}(x,y) > 0$ by the summation over all $(x,y) \in \mathcal{X}\times\mathcal{Y}$. Inequality then follows by the monotonicity of the logarithm function.

For the first inequality, equality holds if and only if $P_{XY}(x,y) = P_X(x)P_Y(y)$ for all $(x,y)$ with $P_{XY}(x,y) > 0$, and for the second inequality, equality holds if and only if $P_{XY}(x,y) = 0$ implies $P_X(x)P_Y(y) = 0$ for any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. It follows that $H(X|Y) = H(X)$ if and only if $P_{XY}(x,y) = P_X(x)P_Y(y)$ for all $(x,y) \in \mathcal{X}\times\mathcal{Y}$. ■

**Proposition 1.5.2 — Chain Rule.** Let $X$ and $Y$ be random variables. Then

$$H(XY) = H(X) + H(Y|X)\,.$$

*Proof.* The chain rule is a simple matter of rewriting:

$$\begin{aligned} H(XY) &= -\sum_{x,y} P_{XY}(x,y)\log P_{XY}(x,y) \\ &= -\sum_{x,y} P_{XY}(x,y)\log\big(P_X(x)P_{Y|X}(y|x)\big) \\ &= -\sum_{x,y} P_{XY}(x,y)\log P_X(x) - \sum_{x,y} P_{XY}(x,y)\log P_{Y|X}(y|x) \\ &= -\sum_{x} P_X(x)\log P_X(x) - \sum_{x} P_X(x)\sum_{y} P_{Y|X}(y|x)\log P_{Y|X}(y|x) \\ &= H(X) + H(Y|X)\,. \end{aligned} \tag{1.19}$$

This was to be shown. ■

The following inequality, also known as the 'independence bound', follows from the fact that $H(Y|X) \le H(Y)$:

**Corollary 1.5.3 — Subadditivity.**

$$H(XY) \le H(X) + H(Y)\,.$$

Equality holds iff $X$ and $Y$ are independent.

Note that applying Definition 1.5.1 to the conditional distribution $P_{XY|\mathcal{A}}$ naturally defines $H(X|Y,\mathcal{A})$, the entropy of $X$ given $Y$ and conditioned on the event $\mathcal{A}$. Since the entropy is a function of the distribution of a random variable, the chain rule also holds when conditioning on an event $\mathcal{A}$. Furthermore, it holds that

$$H(X|YZ) = \sum_{z} P_Z(z) H(X|Y, Z{=}z)\,, \tag{1.20}$$

which is straightforward to verify. With this observation, it is easy to see that the chain rule generalizes as follows.

**Corollary 1.5.4** Let $X$, $Y$ and $Z$ be random variables. Then

$$H(XY|Z) = H(X|Z) + H(Y|XZ)\,.$$

Inductively applying the (generalized) chain rule implies that for any sequence $X_1,\ldots,X_n$ of random variables:

$$H(X_1\cdots X_n) = H(X_1) + H(X_2|X_1) + \cdots + H(X_n|X_{n-1}\cdots X_1)\,. \tag{1.21}$$

**Example 1.5** Consider the binary random variables $X$ and $Y$, with joint distribution

$$P_{XY}(00) = \frac{1}{2}, \quad P_{XY}(01) = \frac{1}{4}, \quad P_{XY}(10) = 0, \quad P_{XY}(11) = \frac{1}{4}.$$

By marginalization, we find that $P_X(0) = \frac{3}{4}$ and $P_X(1) = \frac{1}{4}$, while $P_Y(0) = P_Y(1) = \frac{1}{2}$. This allows us to make the following computations:

$$H(XY) = \frac{1}{2}\log 2 + \frac{1}{4}\log 4 + \frac{1}{4}\log 4 = \frac{3}{2} \tag{1.22}$$

$$H(X) = h\left(\frac{1}{4}\right) = h\left(\frac{3}{4}\right) \approx 0.81 \tag{1.23}$$

$$H(Y) = h\left(\frac{1}{2}\right) = 1 \tag{1.24}$$

$$H(X|Y) = H(XY) - H(Y) = \frac{1}{2} \tag{1.25}$$

$$H(Y|X) = H(XY) - H(X) \approx 0.69 \tag{1.26}$$

We also could have computed $H(X|Y)$ and $H(Y|X)$ directly through the definition of conditional entropy.

Note that for this specific distribution, $H(X|Y=1) > H(X)$. It is important to remember that Proposition 1.5.1 only holds on average, not for specific values of $Y$. Note also that in this example, $H(X|Y) \neq H(Y|X)$.

## 1.6 Mutual Information

**Definition 1.6.1 — Mutual information.** Let $X$ and $Y$ be random variables. The mutual information $I(X;Y)$ of $X$ and $Y$ is defined as

$$I(X;Y) = H(X) - H(X|Y).$$

Thus, in a sense, mutual information reflects the reduction in uncertainty about $X$ when given $Y$. Note the following properties of the mutual information:

$$I(X;Y) = H(X) + H(Y) - H(XY) \qquad \text{(by chain rule)} \tag{1.27}$$

$$I(X;Y) = I(Y;X) \qquad \text{(“symmetry")} \tag{1.28}$$

$$I(X;Y) \geq 0 \qquad \text{(by subadditivity)} \tag{1.29}$$

$$I(X;Y) = 0 \text{ iff } X \text{ and } Y \text{ are independent} \tag{1.30}$$

$$I(X;X) = H(X) \qquad \text{(“self-information”)} \tag{1.31}$$

Applying Definition 1.6.1 to the conditional distribution $P_{XY|\mathcal{A}}$ naturally defines $I(X;Y|\mathcal{A})$, the mutual information of $X$ and $Y$ conditioned on the event $\mathcal{A}$.

**Definition 1.6.2 — Conditional mutual information.** Let $X,Y,Z$ be random variables. Then the conditional mutual information of $X$ and $Y$ given $Z$ is defined as

$$I(X;Y|Z) = \sum_{z} P_Z(z) I(X;Y|Z{=}z),$$

with the convention that the corresponding argument in the summation is 0 for $z$ with $P_Z(z) = 0$.

Conditional mutual information has properties similar to the ones we saw above:

$$I(X;Y|Z) = I(Y;X|Z) \tag{1.32}$$

$$I(X;Y|Z) \geq 0 \tag{1.33}$$

$$I(X;Y|Z) = 0 \text{ iff } X \text{ and } Y \text{ are independent given } Z \tag{1.34}$$

Furthermore, the previous bounds $H(X) \geq 0$, $H(X|Y) \geq 0$, and $I(X;Y) \geq 0$, can all be seen as special cases of $I(X;Y|Z) \geq 0$. These bounds, and any bound they imply, are called **Shannon inequalities**.

It is important to realize that $I(X;Y|Z)$ may be larger or smaller than (or equal to) $I(X;Y)$. The following is easy to verify (and is sometimes used as definition of $I(X;Y|Z)$).

**Proposition 1.6.1** Let $X,Y,Z$ be random variables. Then

$$I(X;Y|Z) = H(X|Z) - H(X|YZ).$$

By this result, we obtain:

**Corollary 1.6.2 — Chain rule for mutual information.** Let $W,X,Y$ and $Z$ be random variables. Then

$$I(WX;Y|Z) = I(X;Y|Z) + I(W;Y|ZX).$$

*Proof.* The proof is a matter of writing out definitions and applying the generalized chain rule.

$$\begin{aligned} I(WX;Y|Z) &= H(WX|Z) - H(WX|YZ) \\ &= (H(X|Z) + H(W|XZ)) - (H(X|YZ) + H(W|XYZ)) \\ &= H(X|Z) - H(X|YZ) + H(W|XZ) - H(W|XYZ) \\ &= I(X;Y|Z) + I(W;Y|XZ). \end{aligned} \tag{1.35}$$

■

## 1.7 Relative entropy

A measure that is related to the mutual information is the relative entropy: it reflects how different two distributions are:

**Definition 1.7.1 — Relative entropy.** The relative entropy (or: **Kullback leibner distance**) of two probability distributions $P$ and $Q$ over the same $\mathcal{X}$ is defined by

$$D(P||Q) := \sum_{\substack{x \in \mathcal{X} \\ P(x)>0}} P(x) \log \frac{P(x)}{Q(x)},$$

where by convention, $\log \frac{p}{0} = \infty$ for all $p$.

Note that if $Q(x) = 0$ for some $x$ with $P(x) > 0$, then $D(P||Q) = \infty$.

**Exercise 1.3** Show that $I(X;Y) = D(P_{XY}||P_X \cdot P_Y)$. ■

This exercise, combined with the equality condition in Theorem 1.7.1 below, shows that the mutual information is a measure of 'how independent' the variables $X$ and $Y$ are: if $P_{XY} = P_X \cdot P_Y$, the variables are independent and their mutual information is zero.

**Theorem 1.7.1 — Information inequality.** For any two probability distributions $P$ and $Q$ defined on the same $\mathcal{X}$,

$$D(P||Q) \geq 0.$$

Equality holds if and only if $P = Q$.

*Proof.* Left as an exercise. Hint: use Jensen's inequality. ■

## 1.8 Entropy Diagrams

We finish this chapter by visually summing up the relations between entropy, joint entropy, conditional entropy, mutual information, and conditional mutual information. For two and three random variables, the relations between these different information-theoretic measures can be nicely represented by means of a Venn-diagram-like **entropy diagram**. The case of two random variables is illustrated in Figure 1.2 (left). From the diagram, one can for instance easily read off the relations $H(X|Y) \leq H(X)$, $I(X;Y) = H(X) + H(Y) - H(XY)$ etc. The case of three random variables is illustrated in Figure 1.2 (right). Also here, one can easily read off all the relations between the information-theoretic measures, like for instance $H(X|YZ) = H(X) - I(X;Z) - I(X;Y|Z)$, which is a relation that is otherwise maybe not immediately obvious.

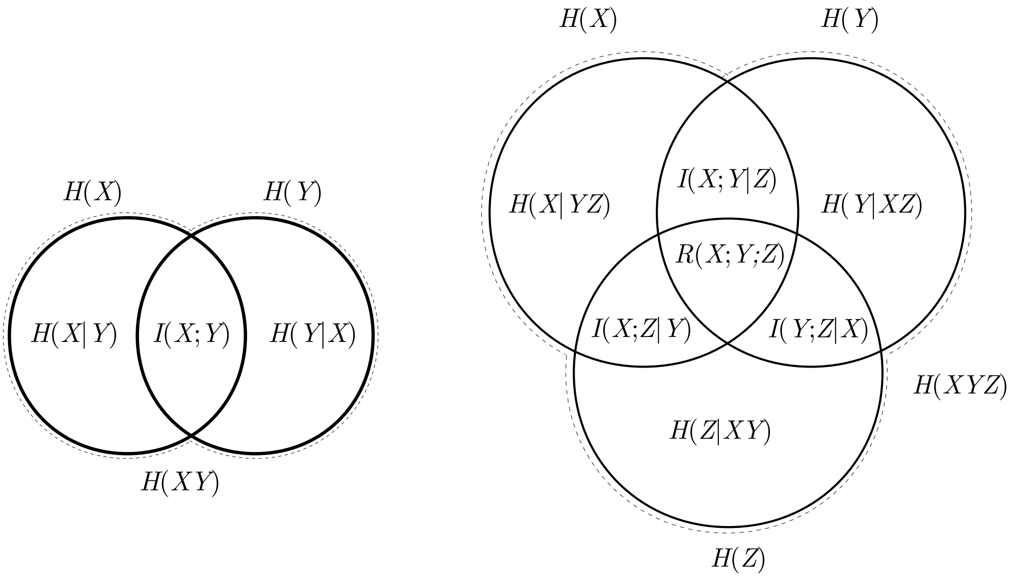


Figure 1.2: Entropy diagram for two (left) and three (right) random variables. The areas encompassed by the dotted lines represent $H(XY)$ and $H(XYZ)$, respectively.

One subtlety with the entropy diagram for three random variables is that the "area in the middle", $R(X;Y;Z) = I(X;Y) - I(X;Y|Z)$, may be *negative*.

## 1.9 Further Reading

- Sections 2.1, 2.2, 3.1-3.3 of [CF]
- Sections 2.1, 2.2, 2.6 of [CT]
- For more background on probability theory, check for instance the lecture script of the Master of Logic course "Basic Probability:Theory" by Philip Schulz and Christian Schaffner.