2 Information Theory

Information theory grew largely out of work published in the late 1940s by Claude Shannon, and stems from a theoretical framework in which stochastic trials represent *communication*, aka *data transmission*. Shannon's work itself can be understood as motivated to a large degree by his cryptographic work at Bletchley Park together with Alan Turing during the Second World War; thus Shannon's model is expressed in terms of *encoding* or *compression*.

2.1 Entropy

- **History:**
  - Terminology from physics (thermodynamics)
  - Entropy rises as energy (heat) is added to a system.
- **Intuitive Definition:**
  - *Entropy* = "chaos", disorder, unpredictability, ...
  - Entropy as a measure of *uncertainty* with respect to the outcome of a stochastic trial:
    * Low entropy → low uncertainty
    * High entropy → high uncertainty

Definition 1 (Entropy) Let X be a random variable with distribution p. Then, the *entropy of* X is written $\mathrm{H}(X)$, and is defined as the mean negative binary logarithm of the probability:

$$\begin{aligned}\mathrm{H}(X) &:= -\sum_{x\in\Omega_X} p(x)\log_2 p(x)\\ &= \sum_{x\in\Omega_X} p(x)\log_2\frac{1}{p(x)}\end{aligned}$$

- *Binary logarithm* used to measure entropy in *bits*: unless otherwise specified, $\log x = \log_2 x$
- By convention, we let $0\log 0 = 0$ and $p\log\frac{p}{0} = \infty$ when computing entropy (and other related quantities).
- Notational variants: $\mathrm{H}(X) = \mathrm{H}(p) = \mathrm{H}_X(p) = \mathrm{H}(p_X)$

Example 1 (Entropy: Fair Coin)

$$\begin{aligned}p(0) &= p(1) = 0.5\\ \mathrm{H}(p) &= -(0.5\log 0.5 + 0.5\log 0.5) = -2\cdot(0.5\cdot -1) = 1\end{aligned}$$

Example 2 (Entropy: Fair Die)

$$
\begin{aligned}
p(i) &= \frac{1}{6} \qquad \text{for } 1 \leq i \leq 6 \\
\mathrm{H}(p) &= 6 \cdot \left(\frac{1}{6} \log 6 \right) = \log 6 \approx 2.58
\end{aligned}
$$

Example 3 (Entropy: Unfair Coin)

$$
\begin{aligned}
p(1) &= 0.2 \qquad , \qquad p(0) = 0.8 \\
\mathrm{H}(p) &= -(0.2 \log 0.2 + 0.8 \log 0.8) \approx 0.722
\end{aligned}
$$

Some Properties of Entropy

- Entropy is non-negative: $\mathrm{H}(X) \geq 0$
- If $p(x) = 1$ for some $x \in \Omega_X$, then $\mathrm{H}(p) = 0$
- There is no *global* upper bound on entropy, but:
  - Uniform distributions maximize entropy (are most unpredictable)
  - In a uniform distribution, all outcomes are equiprobable: $p(x) = \frac{1}{|\Omega_X|}$, so $\mathrm{H}(X) \leq \log |\Omega_X|$.

2.1.1 Entropy and Encoding

Shannon's notion of entropy represents the *average number of bits* required to encode the outcome of a single stochastic trial properly modelled by the distribution p in an *optimal encoding* (read: "maximally compressed"):

- **Recall:** a binary string with n bits can take 2^n possible values – this is just the number of *binary decisions* one has to make to determine which of n *equiprobable events* has occurred (binary search).
- **Idea:** using variable-length codes, an optimal encoding scheme will be one in which *common* messages (read: "outcomes with high probability") are encoded with *fewer bits* than *uncommon* messages.[1]
- **Method:** Knowledge of the probability distribution p gives us a way to determine the minimal number of bits required to encode the occurrence of each outcome x: $\min(\mathrm{length}(\mathrm{code}(x))) = -\log_2 p(x)$; Shannon entropy is just the mean of this quantity.

[1] There's nothing magical about bits here — we could use logarithms of any arbitrary base b to express code lengths in a b-adic number system. Use of the binary (base-2) number system is just a useful convention.

Example 4 (Entropy: DNA) Suppose we wish to encode a particular DNA (sub)sequence; then:

- **Outcomes:**
$$\Omega = \{A, C, T, G\}$$
- **Naïve Code** code_1:
$$\mathrm{A} : 00,\ \mathrm{C} : 01, \mathrm{T} : 10, \mathrm{G} : 11$$
- **Mean (naïve) Code Length:**
$$\begin{aligned} \mathrm{E}(\mathrm{length}(\mathrm{code}_1(X))) &= \sum_{x\in\Omega} p(x) \cdot \mathrm{length}(\mathrm{code}_1(x)) \\ &= (0.5 \cdot 2) + (0.25 \cdot 2) + 2(0.125 \cdot 2) \\ &= 2 \text{ bits} \end{aligned}$$
- **Distribution:**
$$p(A) = 0.5,\ p(C) = 0.25,\ p(T) = 0.125,\ p(G) = 0.125$$
- **Minimal Code Lengths** $= -\log p(x)$:
$$\mathrm{A} : 1 \text{ bit},\ \mathrm{C} : 2 \text{ bits},\ \mathrm{T} : 3 \text{ bits},\ \mathrm{G} : 3 \text{ bits}$$
- **Entropy = Weighted Mean (minimal) Code Length:**
$$\begin{aligned} \mathrm{H}(X) &= \sum_{x\in\Omega} p(x) \cdot \min(\mathrm{length}(\mathrm{code}(x))) \\ &= \sum_{x\in\Omega} p(x) \cdot -\log p(x) \\ &= (0.5 \cdot 1) + (0.25 \cdot 2) + (0.125 \cdot 3) + (0.125 \cdot 3) \\ &= 1.75 \text{ bits} \end{aligned}$$
- **Oops!** Naïve code ain't so great:
$$\mathrm{E}(\mathrm{length}(\mathrm{code}_1(X))) > \mathrm{H}(X)$$
- **Improved Code** code_2:
$$\mathrm{A} : 1,\ \mathrm{C} : 01, \mathrm{T} : 000, \mathrm{G} : 001$$
- **Weighted Mean (improved) Code Length:**
$$\begin{aligned} \mathrm{E}(\mathrm{length}(\mathrm{code}_2(X))) &= \sum_{x\in\Omega} p(x) \cdot \mathrm{length}(\mathrm{code}_2(x)) \\ &= (0.5 \cdot 1) + (0.25 \cdot 2) + 2(0.125 \cdot 3) \\ &= 1.75 \text{ bits} \end{aligned}$$
- **Yipee!** Improved code is optimal.
$$\mathrm{E}(\mathrm{length}(\mathrm{code}_2(X))) = \mathrm{H}(X)$$

2.1.2 Perplexity

- **Idea:** Comparison of information content between two random variables whose sample spaces are of different size – i.e. where we can't simply normalize by $|\Omega|$.
- **Method:** Formalize notion of "information content" in terms of stochastic experiments with uniform distributions, where the only relevant variable is $|\Omega|$.
- **Side Effect:** Perplexity values are much larger than (normalized) entropies.

Definition 2 (Perplexity) For a random variable X with distribution p, the *perplexity* of X is written $\mathrm{G}(X)$ and is defined as:

$$\mathrm{G}(X) := 2^{\mathrm{H}(X)}$$

Intuitively, $G(X) = k$ means that X is just as (un)predictable as a stochastic experiment with k equiprobable possible outcomes.

2.1.3 Joint and Conditional Entropy

- **Idea:** Consider combinations of 2 random variables X and Y.
- **Joint Entropy:** measures unpredictability of *value pairs* $(x, y) \in \Omega_X \times \Omega_Y$.
- **Conditional Entropy:** measures (possibly reduced) unpredictability of an event given knowledge of a (different) event – allows us to quantify dependence.

Definition 3 (Joint Entropy) For two random variables X and Y, the *joint entropy* of X and Y is written $\mathrm{H}(X, Y)$ and is defined as:

$$\mathrm{H}(X, Y) = -\sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} p(x, y) \cdot \log p(x, y)$$

Definition 4 (Conditional Entropy) For two random variables X and Y, the *conditional entropy* of Y given X is written $\mathrm{H}(Y|X)$ and is defined as:

$$\begin{aligned} \mathrm{H}(Y|X) &= \sum_{x \in \Omega_X} p(x) \mathrm{H}(Y|X = x) \\ &= \sum_{x \in \Omega_X} p(x) \left(- \sum_{y \in \Omega_Y} p(y|x) \log p(y|x) \right) \end{aligned}$$

$$
\begin{aligned}
&= -\sum_{x\in\Omega_X}\sum_{y\in\Omega_Y} p(x)p(y|x)\log p(y|x) \\
&= -\sum_{x\in\Omega_X}\sum_{y\in\Omega_Y} p(x,y)\cdot\log p(y|x) \\
&= \mathrm{H}(X,Y) - \mathrm{H}(X)
\end{aligned}
$$

Some Properties of Joint and Conditional Entropy

- **Chain Rule**

$$\mathrm{H}(X,Y) \;=\; \mathrm{H}(X|Y) + \mathrm{H}(Y) = \mathrm{H}(Y|X) + \mathrm{H}(X)$$

- **Conditional Entropy Maximum**

$$\mathrm{H}(Y|X) \leq \mathrm{H}(Y)$$

- **Addition Rule**

$$\mathrm{H}(X,Y) \leq \mathrm{H}(X) + \mathrm{H}(Y)$$

  If X and Y are independent, then:

$$\mathrm{H}(X,Y) = \mathrm{H}(X) + \mathrm{H}(Y)$$

2.2 Relative Entropy

- **Idea:** measure similarity between two distributions p and q.
- **Method:** compute mean number of bits wasted when encoding events governed by one distribution with an optimal code for the other.

Definition 5 (Relative Entropy) Let p and q be probability distributions over a set Ω of basic outcomes. The *relative entropy of p and q* — also known as the *Kullback-Leibler divergence of p and q* — is written $\mathrm{D}(p\|q)$, and defined as the average number of bits wasted when encoding a stochastic process with distribution p under an optimal code for q:

$$
\begin{aligned}
\mathrm{D}(p\|q) &= \sum_{x\in\Omega} p(x)\log\frac{p(x)}{q(x)} \\
&= E_p\left(\log\frac{p(x)}{q(x)}\right)
\end{aligned}
$$

Some Properties of Relative Entropy

- Relative entropy is always non-negative: $\mathrm{D}(p\|q) \geq 0$.

- $\mathrm{D}(p\|q) = 0$ iff $p = q$
- **Caveats:** Relative entropy is **not** a *metric*:
  - No Symmetry: $\diamond\mathrm{D}(p\|q) \neq \mathrm{D}(q\|p)$
  - No Triangle Inequality: $\diamond\mathrm{D}(p\|q) \not\leq \mathrm{D}(q\|p)$

2.3 Mutual Information

- **Idea:** Exploit dependence when simultaneously encoding outcomes of two stochastic processes.
- **Method:** Compute relative entropy between the actual joint distribution and an independent distribution – essentially an *information gain* ratio with respect to the assumption that the two distributions are independent.

Definition 6 (Mutual Information) Let X and Y be random variables. The *mutual information* between X and Y is written $I(X;Y)$ and defined as the relative entropy of the joint distribution and an independent distribution:

$$\begin{aligned} I(X;Y) &= \mathrm{D}(p(X,Y)\|p(X)p(Y)) \\ &= E_{p(X,Y)}\left(\log \frac{p(X,Y)}{p(X)p(Y)}\right) \\ &= \sum_{x\in\Omega_X}\sum_{y\in\Omega_Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \end{aligned}$$

Some Properties of Mutual Information

- **Relation to Entropy:**

$$\begin{aligned} I(X;Y) &= \mathrm{H}(X) - \mathrm{H}(X|Y) \\ &= \mathrm{H}(Y) - \mathrm{H}(Y|X) \\ &= \mathrm{H}(X) + \mathrm{H}(Y) - \mathrm{H}(X,Y) \\ &= I(Y;X) \\ I(X;X) &= \mathrm{H}(X) \end{aligned}$$

Definition 7 (Pointwise Mutual Information) Let x and y be values of random variables X and Y, respectively: $x \in \Omega_X, y \in \Omega_y$. The *pointwise mutual information* between x and y is written $I(x,y)$ and defined:

$$I(x,y) = \log\frac{p(x,y)}{p(x)p(y)}$$

Pointwise MI is symmetric, but may be negative. It can be used as an indicator of the association between individual elements (points) x and y, but is highly sensitive to low probabilities, so it is sometimes additionally weighted by e.g. $p(x,y)$.

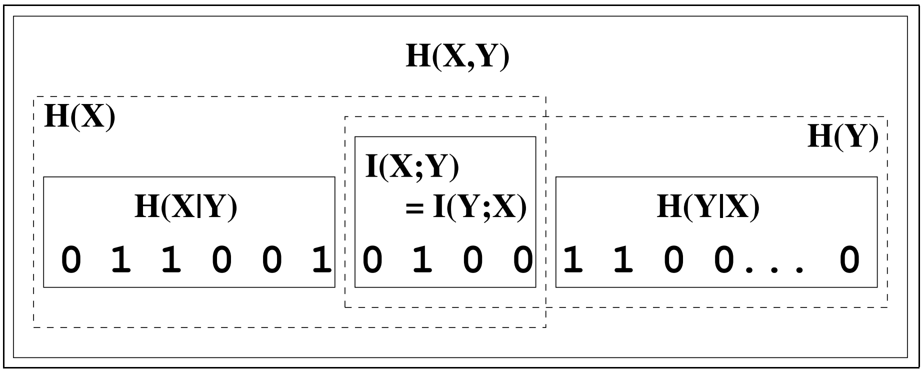

Figure 1: Mutual Information and various entropies