# Lecture: Definition Entropy and Mutual information

**Theorem** (KL1)**. *Positivity of KL*** *[1, Th. 2.6.3]:*

$$D(p||q) \geq 0 \tag{1}$$

*with equality iff* $\forall x, p(x) = q(x)$.

This is a consequence of Jensen's inequality [1, Th. 2.6.2]:
If $f$ is a convex function and $Y$ is a random variable with numerical values, then

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

with equality when $f(.)$ is not strictly convex, or when $f(.)$ is strictly convex and $Y$ follows a degenerate distribution (i.e. is a constant).

*Proof.* Let $X \sim p(x)$. Let $q(x)$ be another distribution defined on the same alphabet $\mathcal{X}$. Let $Supp(p) = \{x : p(x) > 0\}$.
Let $Y = \dfrac{q(X)}{p(X)}$, where $p(X) > 0$. $Y$ is a r.v. with numerical values. More precisely,

$$Y = \begin{cases} \dfrac{q(x)}{p(x)} & \text{with probability } p(x), \text{ if } p(x) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

$$\begin{aligned} D(p||q) &= \mathbb{E}_X \left[\log_2 \frac{p(X)}{q(X)}\right] && (3)\\ &= -\mathbb{E}_Y \left[\log_2 Y\right] && (4)\\ &\geq -\log_2 \mathbb{E}_Y \left[Y\right] && (5)\\ &\geq 0 && (6) \end{aligned}$$

where (5) follows from the convexity of the function $-\log_2(.)$ and Jensen's inequality, and where (6) follows from

$$\mathbb{E}[Y] = \sum_{x \in Supp(p)} p(x) \frac{q(x)}{p(x)} \tag{7}$$

**Case of equality in** (1). From the case of equality in Jensen's inequality (see [1, Th. 2.6.2] and reminder above), and from the fact that $-\log$ is strictly convex, there is equality in (1) if $Y$ is deterministic i.e. $\frac{q(X)}{p(X)}$ is constant a.s. i.e.

$$\forall x \in Supp(p), q(x) = cp(x), \text{ where } c \in \mathbb{R}. \tag{8}$$

Moreover, there is equality in (6), therefore, from (7), we have that

$$\sum_{x \in Supp(p)} q(x) = 1. \tag{9}$$

Therefore $\sum_{x \in Supp(p)} q(x) = 1 = c \sum_{x \in Supp(p)} p(x)$, which leads to $c = 1$. i.e.

$$\begin{cases} \forall x \in Supp(p), q(x) = p(x) \\ \forall x \in \mathcal{X}/Supp(p), q(x) = p(x) = 0. \end{cases} \tag{10}$$

Conversely if (10), $D(p||q) = 0$.

□

# Lecture: Variable Length Coding - Zero error data compression

**Theorem 1** (uniquely decodable code $\Leftrightarrow$ KI)**.** *[1, Th 5.5.1]*
*The codeword lengths of any uniquely decodable code (UDC) must satisfy the Kraft inequality (KI):*

$$\sum_{i=1}^{M} D^{-l_i} \leq 1 \tag{11}$$

***Conversely,*** *given a set of codeword lengths that satisfy this inequality, it is possible to construct a uniquely decodable code with these codeword lengths.*

*Proof.* **Sufficient condition**: KI (11) $\Rightarrow$ UDC.
It was shown that (11) $\Rightarrow \exists$ a prefix code. So in particular a UDC.

**Necessary condition**: UDC $\Rightarrow$ KI (11).
Let $C$ be a UDC. Let $l_{\max} = \max_i l_i$.

$$\sum_{i=1}^{M} D^{-l_i} = \sum_{j=1}^{l_{\max}} w_j D^{-j}, \quad w_j = \# \text{ of codewords of length } j \tag{12}$$

$$\left(\sum_{j=1}^{l_{\max}} w_j D^{-j}\right)^n = \underbrace{\sum_{j_1} \dots \sum_{j_n}}_{1 \leq j_k \leq l_{\max}} w_{j_1} \dots w_{j_n} D^{-j_1} \dots D^{-j_n} \tag{13}$$

$$= \sum_{k=n}^{nl_{\max}} N_k D^{-k} \tag{14}$$

since $\forall k, 1 \leq j_k \leq l_{\max} \Rightarrow n \leq j_1 + \ldots + j_n \leq nl_{\max}$, and where $N_k$ is the number of sequence of codewords of length $k$, corresponding to the encoding of $n$ source symbols.

But UDC implies that 2 different source messages have 2 different codeword sequences. Therefore the $N_k$ codewords are distincts.
Therefore, $N_k \leq D^k = \#$ of possible sequences with $k$ letters. Therefore,

$$\left(\sum_{j=1}^{l_{\max}} w_j D^{-j}\right)^n = \sum_{k=1}^{nl_{\max}} N_k D^{-k} \leq \sum_{k=n}^{nl_{\max}} D^k D^{-k} \leq nl_{\max} \tag{15}$$

$$\sum_{j=1}^{l_{\max}} w_j D^{-j} \leq n^{\frac{1}{n}} l_{\max}^{\frac{1}{n}} = e^{\frac{1}{n}\log(nl_{\max})} \xrightarrow[n\to\infty]{} e^0 = 1 \tag{16}$$

Therefore,

$$\sum_{j=1}^{l_{\max}} w_j D^{-j} \leq 1 \tag{17}$$

□

**Theorem 2** (Expected length of a Shannon code [CT Sec. 5.4])**.** *Let $X$ be a r.v. with entropy $H(X)$. The* ***Shannon code*** *for the source $X$ can be turned* ***into a prefix code*** *and its* ***expected length*** *$L(C)$* ***satisfies***

$$\frac{H(X)}{\log D} \leq L(C) < \frac{H(X)}{\log D} + 1 \tag{18}$$

*Proof.* For the $i^{th}$ symbol of the alphabet of $X$ with probability $p_i > 0$, the Shannon code assign a codeword of length $l_i = \lceil -\log_D(p_i) \rceil \Leftrightarrow -\log_D(p_i) \leq l_i < -\log_D(p_i) + 1$.

- First, the set of lengths $\{l_i\}_i$ satisfies the Kraft inequality. Indeed, since $p_i > 0$ (we encode only symbols with non zero probability)

$$-\log_D(p_i) \leq l_i \Leftrightarrow D^{-l_i} \leq p_i \Rightarrow \sum_i D^{-l_i} \leq 1$$

  Therefore, there exists a prefix code with the codeword lengths of the Shannon code i.e. any Shannon code can be turned into a prefix code.
- The expected length of the Shannon code satisfies

$$-\sum_i p_i \log_D(p_i) \leq \sum_i p_i l_i < -\sum_i p_i \log_D(p_i) + 1 \Leftrightarrow \frac{H(X)}{\log D} \leq L(C) < \frac{H(X)}{\log D} + 1$$

□

**Theorem 3** (Lower and upper bound on the expected length of an optimal code [CT 5.4.1])**.** *Let $X$ be a r.v. with entropy $H(X)$. Any* ***optimal code*** *$C^*$ for $X$ with codeword lengths $l_1^*, ..., l_M^*$ and* ***expected length $L(C^*) = \sum p_i l_i^*$ satisfies***

$$\frac{H(X)}{\log D} \leq L(C^*) < \frac{H(X)}{\log D} + 1$$

*Proof.*

- Upper bound: The code is optimal so it is better than a Shannon code $C$:

$$L(C^*) \leq L(C) < \frac{H(X)}{\log D} + 1$$

- Lower bound: Any prefix code satisfies the lower bound. So does the optimal.

$$\frac{H(X)}{\log D} \leq L(C^*)$$

□

**Lemma 1** (Necessary conditions on optimal prefix codes[CT Le5.8.1])**.** *Given a* ***binary*** *prefix code $C$ with word lengths $l_1, ..., l_M$ associated with a set of symbols with probabilities $p_1, ..., p_M$.*
*Without loss of generality, assume that*
*(i) $p_1 \geq p_2 \geq ... \geq p_M$,*
*(ii) a group of symbols with the same probability is arranged in order of increasing codeword length (i.e. if $p_i = p_{i+1} = ... = p_{i+r}$ then $l_i \leq l_{i+1} ... \leq l_{i+r}$).*
*If* ***$C$ is optimal*** *within the class of* ***prefix*** *codes,* ***$C$ must satisfy****:*

1. ***higher*** *probabilities symbols have* ***shorter*** *codewords ($p_i > p_k \Rightarrow l_i < l_k$),*
2. *the two least probable symbols have* ***equal*** *length ($l_M = l_{M-1}$),*
3. *among the codewords of* ***length $l_M$****, there must be at least two words that* ***agree in all digits except the last****.*

*Proof.*

1. By contradiction. Assume $p_i > p_k$ and $l_i > l_k$. Let $C'$ be the code where we exchange the codewords of $i$ and $k$.

$$L(C) - L(C') = p_i l_i + p_k l_k - p_i l_k - p_k l_k = \underbrace{(p_i - p_k)}_{>0} \underbrace{(l_i - l_k)}_{>0} > 0$$

   Therefore there exists a code $C'$ with shortest expected length than $C$. This contradicts the fact that $C$ is optimal.

2. By contradiction. Assume $l_M \neq l_{M-1}$.

- Then, necessarily $l_M < l_{M-1}$. Indeed, $l_M \leq l_{M-1}$ holds in general either due to Condition 1, if $p_{M-1} > p_M$, or due to condition (ii) if $p_{M-1} > p_M$.
- If $l_M < l_{M-1}$, then the codeword for the $M-1^{th}$ symbol is not prefix of the codeword for symbol $M$
  i.e. the $M-1$ first letters of the codeword for the $M-1^{th}$ symbol differ from the $M-1$ first letters of the codeword for symbol $M$
  Therefore we can eliminate without any ambiguity the last letter of the codeword for $M$.
  This builds a new code $C'$, which is prefix and is shorter than $C$. This contradicts the fact that $C$ is optimal.

3. By contradiction. Condition 3. states that if symbols $\alpha_{M-1}$ and $\alpha_M$ have associated codewords of length $l_M$, then these symbols admit the following codewords:

$$\alpha_{M-1} : \xleftrightarrow{\quad l_M - 1 \quad} 0$$
$$\alpha_M : \xleftrightarrow{\quad l_M - 1 \quad} 1$$

If Condition 3 is not satisfied, then symbols $\alpha_{M-1}$ and $\alpha_M$ admit the following codewords:

$$\alpha_{M-1} : \xleftrightarrow{\quad l_M - 2 \quad} 0\ 1$$
$$\alpha_M : \xleftrightarrow{\quad l_M - 2 \quad} 1\ 0$$

Then, we can eliminate without any ambiguity the last letter of the codeword for symbols $\alpha_{M-1}$ and $\alpha_M$.
This builds a new code $C'$, which is prefix and is shorter than $C$. This contradicts the fact that $C$ is optimal.

□

**Theorem 4** (Huffman code is optimal [CT Th. 5.8.1])**.** *If $C$ is a Huffman code and $C'$ is any other uniquely decodable code, $L(C) \leq L(C')$.*

*Proof.* By contradiction. Let us assume that $C_1$ is not optimal i.e. $\exists C'_1$ for $\{\alpha_1, \alpha_2, ..., \alpha_M\}$ with codewords $W'_1, ..., W'_M$ of length $l'_1, ..., l'_M$ ans $C'1$ is optimal. *Let us show now that this contradicts the fact that $C_2$ is optimal.*

$$\begin{aligned} C'1 \text{ is optimal } &\Rightarrow l'_M = l'_{M-1}. \text{ (condition 2)} \\ &\Rightarrow \text{ at least two codewords agree in all digits except the last. (condition 3)} \\ &\quad\ \text{Let us assume that this is the case for } W_{M-1}, W_M \end{aligned}$$

Let us construct $C'_2$ from $C'_1$ for symbols $\{\alpha_1, ..., \alpha_{M-2}, \alpha_{M-1,M}\}$, where ,$\alpha_{M-1,M}$ is the combination of $\alpha_{M-1}$ and $\alpha_M$ s.t.

$$\alpha_i \mapsto W'_i$$
$$\alpha_{M-1,M} \mapsto W'_{M-1,M}$$

where $W'_{M-1,M}$ contains the first $l'_M - 1$ bits except the last one of $W_{M-1}$ and $W_M$ (this is possible see above the condition 3).
Finally, we get

$$\begin{aligned} L(C'_2) - L(C_2) &= \sum_{i=1}^{M-2} p_i l'_i + (p_M + p_{M-1})(l'_M - 1) - \sum_{i=1}^{M-2} p_i l_i - (p_M + p_{M-1})(l_M - 1) \\ &= \sum_{i=1}^{M} p_i l'_i - (p_M + p_{M-1}) - \sum_{i=1}^{M} p_i l_i + (p_M + p_{M-1}) \\ &= L(C'_1) - L(C_1) < 0 \end{aligned}$$

since $C_1$ is not optimal and $C'_1$ is. This contradicts the fact that $C_2$ is optimal.

□

## References

[1] T. Cover and J. Thomas, *Elements of information theory, second Edition*. Wiley, 2006.