EL-GY 6063: Information Theory
Lecture 2

February 4, 2019

Definition 1. *Discrete Source: Random variable X characterized by $(\mathcal{X}, 2^{\mathcal{X}}, P_X)$, where $\mathcal{X} = \{x_1, x_2, \cdots, x_{|\mathcal{X}|}\}$ is a subset of the real numbers and:*

$$P_X : \left(P_X(x_1), \quad P_X(x_2), \quad \cdots \quad P_X(x_{|\mathcal{X}|})\right), \qquad P_X(x_i) \geq 0, \forall i \in \{1, 2, \cdots, |\mathcal{X}|\}, \qquad \sum_{i=1}^{|\mathcal{X}|} P_X(x_i) = 1.$$

Example 1. *Let X_B, X_M and X_G be the indicator function of the event that tomorrow is a sunny day in Brooklyn, Manhattan and the Grand Canyon, respectively. X_B, X_N and X_G are three discrete information sources. (The Grand Canyon has an extremely dry climate and is usually sunny, whereas Manhattan and Brooklyn have a climate that undergoes rapid daily changes.)*

Definition 2. *Let X be a discrete source characterized by $(\mathcal{X}, 2^{\mathcal{X}}, P_X)$. The entropy of X is defined as:*

$$H(X) \triangleq -\sum_{x \in \mathcal{X}} P_X(x) \log P_X(x).$$

By convention, we take the base of the logarithm to be equal to 2. In this case the unit of entropy is bits. Sometimes, the base of the logarithm is taken to be e. In this case, the unit of entropy is nats. Furthermore, we define $0 \cdot \log 0 = 0$. The entropy of X is a function of the probability distribution P_X and is often written as $H(P_X(x_1), P_X(x_2), \cdots, P_X(x_{|\mathcal{X}|}))$.

Note that the entropy function can also be expressed as $H(X) = E_x(-\log P_X(x))$, where E_x is the expectation with respect to x.

Example 2. *Assume that through empirical observations, it is known that $P_{X_B}(1) = \frac{1}{8}$ and $P_{X_G}(1) = \frac{19}{20}$. Then,*

$$\begin{aligned} H(X_B) &= -\frac{1}{8}\log_2(\frac{1}{8}) - \frac{7}{8}\log_2(\frac{7}{8}) \approx 0.54 \quad bits, \\ H(X_G) &= -\frac{19}{20}\log_2(\frac{19}{20}) - \frac{1}{20}\log_2(\frac{1}{20}) \approx 0.28 \quad bits. \end{aligned}$$

1 Axiomatic Derivation of Entropy

References:
1- Problem 2.46 Textbook
2- Information Theory Coding Theorems for Discrete Memoryless Systems, Chapter 1 Problems 11-14 (available online)
3- Communication system - Sam Shanmugam - Chapter 4

Let us consider a discrete information source X with $(\mathcal{X}, 2^{\mathcal{X}}, P_X)$ and $\mathcal{X} = \{x_1, x_2, \cdots, |\mathcal{X}|\}$. Let's assume that there is a function $h : \mathcal{X} \to \mathbb{R}^+$, where $h(x_i), i \in \{1, 2, \cdots, |\mathcal{X}|\}$ represents the information revealed by knowing that $X = x_i$. Then, the average information revealed by knowing the value of X is $\tilde{H}(X) \triangleq E(h(X)) \triangleq \tilde{H}(P_X(x_1), P_X(x_2), \cdots, P_X(x_{|\mathcal{X}|}))$. This is also called the amount of uncertainty in X.

Note that $\tilde{H}$ can be viewed as a function from the set of all possible discrete probability distributions to the set of non-negative real numbers. More precisely, let $\mathcal{P}$ be the set of all discrete probability distributions. Then, $\tilde{H} : \mathcal{P} \to R^{\geq 0}$. We will use the following axioms to prove that $\tilde{H}(X)$ is the entropy function multiplied by a positive constant:
Axiom 1: If $|\mathcal{X}| = m$, where m is a natural number and X is uniformly distributed on $\mathcal{X}$ (i.e. $P_X(x_i) = \frac{1}{m}, \forall i \in \{1, 2, \cdots, m\}$), then:

$$\tilde{H}(X) = \tilde{H}(\frac{1}{m}, \frac{1}{m}, \cdots, \frac{1}{m}) \triangleq f(m),$$

The function $f(m)$ is a non-decreasing function of m since increasing the number of outcomes increases the information gained by knowing the outcome for the uniform random variable.
Axiom 2: Assume that X and Y are two uniformly distributed random variables on the sets $\{1, 2, \cdots, m\}$ and $\{1, 2, \cdots, \ell\}$, respectively, then:

$$\tilde{H}(X, Y) = f(mL) = f(m) + f(\ell).$$

Axiom 3: (The grouping axiom) For the random variable X characterized by $(\mathcal{X}, 2^{\mathcal{X}}, P_X)$, let $(\mathcal{A}, \mathcal{B})$ be a partition of $\mathcal{X}$ (i.e. i) $\mathcal{A} \cup \mathcal{B} = \mathcal{X}$, ii) $\mathcal{A} \cap \mathcal{B} = \phi$.). Then, let $\mathbb{1}(\mathcal{A})$ be the indicator of the event that X is in $\mathcal{A}$. Define $X_{\mathcal{A}}$ as the random variable characterized by $(\mathcal{A}, 2^{\mathcal{A}}, P_{X|X\in\mathcal{A}})$ and $X_{\mathcal{B}}$ the random variable characterized by $(\mathcal{B}, 2^{\mathcal{B}}, P_{X|X\in\mathcal{B}})$, then:

$$\tilde{H}(X) = \tilde{H}(\mathbb{1}(\mathcal{A})) + P(X \in \mathcal{A})\tilde{H}(X_{\mathcal{A}}) + P(X \in \mathcal{B})\tilde{H}(X_{\mathcal{B}}).$$

This can be interpreted as follows: The information in X can be expressed as the information revealed by knowing whether $X \in \mathcal{A}$ plus the average of the remaining information given this knowledge of X.

Axiom 4: (Continuity Axiom) If X is a Bernoulli random variable with parameter $p \in (0, 1)$, then $\tilde{H}(X)$ is continuous in p.

Theorem 1. *Let $\mathcal{P}$ be the set of all discrete probability distributions. Then, the function $\tilde{H} : \mathcal{P} \to \mathbb{R}^{\geq 0}$ satisfies the four axioms if and only if:*

$$\tilde{H}(X) = -c \sum_{x \in \mathcal{X}} P_X(x) \cdot \log_2(P_X(x)) = cH(X), \tag{1}$$

where $c \geq 0$ is a constant. Particularly if the entropy of the binary symmetric source (i.e. $P_X(1) = P_X(0) = \frac{1}{2}$) is taken to be equal to one. Then, $\tilde{H}(X) = H(X)$.

Proof. First, note that the function given in Equation (1) satisfies the four axioms (Exercise). Next, we prove that any function satisfying the axioms is in the form of Equation (1).
Step 1: We prove that $\forall n, m \in \mathbb{N} : f(m) = c \log m$ for a non-negative constant c.
From axiom 2 we have:

$$f(m^2) = f(m \cdot m) = f(m) + f(m) = 2f(m),$$

by induction we get $f(m^n) = nf(m)$. Particularly, $f(1) = f(1^2) = 2f(1) \Rightarrow f(1) = 0$ and from axiom 1 we have $f(m) \geq 0, \forall m > 1$ since it is increasing in m.

Next, fix $m \in \mathbb{N}$ and let $r > 0$ be an arbitrary number. From calculus there exists a positive integer k such that:

$$m^k \leq 2^r < m^{k+1}. \tag{2}$$

Note that $\log_2$ is an increasing function, so from (2), we have:

$$\log m^k \leq \log_2 2^r < \log_2 m^{k+1} \Rightarrow k \log_2 m \leq r < (k+1) \log_2 m \Rightarrow \frac{k}{r} \leq \frac{1}{\log_2 m} < \frac{k+1}{r}.$$

Similarly, note that from axiom 1, we have that $f(m)$ is increasing in m. So,

$$f(m^k) \leq f(2^r) < f(m^{k+1}) \Rightarrow kf(m) \leq rf(2) < (k+1)f(m) \Rightarrow \frac{k}{r} \leq \frac{f(2)}{f(m)} < \frac{k+1}{r}.$$

Comparing these two equations we get:

$$|\frac{f(2)}{f(m)} - \frac{1}{\log_2 m}| \leq \frac{1}{r}, \forall r > 0.$$

Take $r \to \infty$ to conclude that $\frac{f(2)}{f(m)} - \frac{1}{\log_2 m} = 0 \Rightarrow f(m) = f(2)\log_2(m) = c\log_2(m)$ where $c = f(2) \geq 0$.
Step 2: We prove that if X is a Bernoulli random variable with parameter $p \in [0,1]$, where p is a rational number, then $H(X) = -c(p\log_2 p + (1-p)\log_2(1-p))$.
Assume that $p = \frac{k}{n}$ where $k, n \in \mathbb{N}$ and $k \leq n$. Define Y as the random variable that is distributed uniformly over the alphabet $\{1, 2, \cdots, n\}$. Let $\mathcal{A} = \{1, 2, \cdots, k\}$. Then using axiom 3 we have,

$$\tilde{H}(Y) = \tilde{H}(\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n}) = \tilde{H}(P_Y(\mathcal{A}), P_Y(\mathcal{A}^c)) + P_Y(\mathcal{A})\tilde{H}(\frac{1}{k}, \frac{1}{k}, \cdots, \frac{1}{k}) + P_Y(\mathcal{A}^c)\tilde{H}(\frac{1}{n-k}, \frac{1}{n-k}, \cdots, \frac{1}{n-k})$$
$$\Rightarrow f(n) = \tilde{H}(X) + \frac{k}{n}f(k) + \frac{n-k}{n}f(n-k).$$

The result follows from step 1.
Step 3: Assume that X is a random variable with distribution $P_X(x_i), i \in \{1, 2, \cdots, |\mathcal{X}|\}$ such that $P_X(x_i)$ are rational numbers. Then,

$$\tilde{H}(X) = -c\sum_{i=1}^{|\mathcal{X}|} P_X(x_i)\log_2 P_X(x_i),$$

where $c > 0$. (Proof is similar to the previous step and is left as an exercise.)
Step 4: Let X be a random variable with distribution $P_X(x_i), i \in \{1, 2, \cdots, |\mathcal{X}|\}$ where $P_X(x_i)$ are real numbers. Then

$$\tilde{H}(X) = -c\sum_{i=1}^{|\mathcal{X}|} P_X(x_i)\log_2 P_X(x_i),$$

This follows from axiom 4 and the fact that rational numbers are dense in real numbers. □

Example 3. *Let X be a binary symmetric source (i.e. $P_X(0) = P_X(1) = \frac{1}{2}$). Then, $H(X) = 1 \quad bit$.*

Proof.

$$H(X) = -(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}) = -\log_2\frac{1}{2} = \log_2 2 = 1.$$

□

Example 4. *Let X be a constant random variable (i.e. $P_X(c) = 1$, for some $c \in \mathbb{R}$.). Then, $H(X) = 0$.*

Proof. $H(X) = -1 \cdot \log_2 1 = 0$. □

Example 5. *Let X be a Bernoulli random variable with parameter $p \in [0,1]$. Then, $H(p, 1-p)$ is increasing for $p < \frac{1}{2}$ and decreasing for $p > \frac{1}{2}$.*

Proof.

$$\frac{\delta}{\delta p}H_b(p) = \frac{\delta}{\delta p}(-p\log_2 p - (1-p)\log_2(1-p) = -\log_2 p + \log_2(1-p) - \frac{1}{\log_e 2}.1 + \frac{1}{\log_e 2} = \log_2\frac{1-p}{p}.$$

Note that:

$$\log_2\frac{1-p}{p} \geq 0 \Rightarrow \frac{1-p}{p} \geq 1 \Rightarrow \frac{1}{2} \geq p.$$

□

2 Properties of Entropy

Next, we will study some of the properties of the entropy function.

Lemma 1 (Gibbs). *Let $(P_1, P_2, \cdots, P_n)$ and $(Q_1, Q_2, \cdots, Q_n)$ be two probability distributions on the alphabet $\{1, 2, \cdots, n\}$. Then,*

$$\sum_{i=1}^{n} P_i \log_2 \frac{Q_i}{P_i} \leq 0, \text{“} = \text{”} \textit{ iff } P_i = Q_i, \forall i.$$

Proof. From calculus we know that $\log_e x \leq x - 1, \forall x > 0$ and “ $=$ ” iff $x = 1$. So, $\log_e \frac{Q_i}{P_i} \leq \frac{Q_i}{P_i} - 1, \forall i$. So,

$$\begin{aligned}
&\sum_{i=1}^{n} P_i \log_2 \frac{Q_i}{P_i} = \frac{1}{\log_e 2} \sum_{i=1}^{n} P_i \log_e \frac{Q_i}{P_i} \leq \frac{1}{\log_e 2} \sum_{i=1}^{n} P_i (\frac{Q_i}{P_i} - 1) = \frac{1}{\log_e 2} \sum_{i=1}^{n} (P_i - Q_i) \\
&= \frac{1}{\log_e 2} (\sum_{i=1}^{n} P_i - \sum_{i=1}^{n} Q_i) = \frac{1}{\log_e 2} (1 - 1) = 0,
\end{aligned}$$

and for equality we need $\frac{Q_i}{P_i} = 1, \forall i$. □

Lemma 2. *Let X be defined on the alphabet $\mathcal{X}$ where $|\mathcal{X}| = m$. Then,*

$$H(X) \leq \log m, \text{“} = \text{”} \textit{ iff } P_i = \frac{1}{m}, \forall i.$$

Proof. In the Gibbs lemma, set $Q_i = \frac{1}{m}, \forall i$. □

Definition 3. *For two discrete random variables X and Y defined on probability spaces $(\mathcal{X}, 2^{\mathcal{X}}, P_X)$ and $(\mathcal{Y}, 2^{\mathcal{Y}}, P_Y)$, respectively, the joint entropy is defined as:*

$$H(X, Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_{X,Y}(x, y) \log_2 P_{X,Y}(x, y).$$

Similarly, for the vector of discrete random variables $X^n = (X_1, X_2, \cdots, X_n)$:

$$H(X_1, X_2, \cdots, X_n) = -\sum_{x^n \in \mathcal{X}^n} P_{X^n}(x^n) \log_2 P_{X^n}(x^n).$$

Example 6. *Let $X = Y$ with probability one. Then $H(X, Y) = H(X) = H(Y)$.*

Proof.

$$\begin{aligned}
H(X, Y) &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_{X,Y}(x, y) \log_2 P_{X,Y}(x, y) \\
&= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_X(x) P_{Y|X}(y|x) \log_2 P_X(x) P_{Y|X}(y|x) \\
&= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_X(x) \mathbb{1}(x = y) \log_2 P_X(x) \mathbb{1}(x = y) \\
&= -\sum_{x \in \mathcal{X}} P_X(x) \log_2 P_X(x) \\
&= H(X).
\end{aligned}$$

□

Lemma 3. *For two random variables X and Y, $H(X, Y) \leq H(X) + H(Y)$ and “ $=$ ” iff X and Y are independent.*

Proof. Note that:

$$H(X) = -\sum_{x\in\mathcal{X}} P_X(x)\log P_X(x) = -\sum_{x,y\in\mathcal{X}\times\mathcal{Y}} P_{X,Y}(x,y)\log P_X(x),$$
$$H(Y) = -\sum_{y\in\mathcal{Y}} P_Y(y)\log P_Y(y) = -\sum_{x,y\in\mathcal{X}\times\mathcal{Y}} P_{X,Y}(x,y)\log P_Y(y),$$

So,

$$\begin{aligned} H(X)+H(Y)-H(X,Y) &= -\sum_{x,y\in\mathcal{X}\times\mathcal{Y}} P_{X,Y}(x,y)\log P_X(x) - \sum_{x,y\in\mathcal{X}\times\mathcal{Y}} P_{X,Y}(x,y)\log P_Y(y) \\ &\quad + \sum_{x,y\in\mathcal{X}\times\mathcal{Y}} P_{X,Y}(x,y)\log P_{X,Y}(x,y) \\ &= \sum_{x,y\in\mathcal{X}\times\mathcal{Y}} P_{X,Y}(x,y)\log\frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)} \geq 0, \end{aligned}$$

where we have used the Gibbs Lemma by setting $Q_{X,Y}(x,y) = P_X(x)P_Y(y)$ in the last inequality and equality holds iff $P_{X,Y}(x,y) = P_X(x)P_Y(y), \forall x, y$. □

Lemma 4. *For the vector of random variables X^n, we have $H(X^n) \leq \sum_{i=1}^n H(X_i)$, and " $=$ " iff X_is are mutually independent.*

The proof is left as an exercise.

3 Conditional Entropy

Let X and Y be two random variables. The conditional entropy of the random variable X given that $Y=y$ is defined as $H(X|Y=y) \triangleq \sum_{x\in\mathcal{X}} P_{X|Y}(x|y)\log P_{X|Y}(x|y)$. This can be interpreted as the amount of uncertainty (information) revealed by knowing the value of X given that we know that $Y=y$.

The conditional entropy of the random variable X given the random variable Y is defined as the average $H(X|Y) \triangleq \sum_{x,y\in\mathcal{X}\times\mathcal{Y}} P_{X,Y}(x,y)\log P_{X|Y}(x|y) = E_y(H(X|Y=y))$. This can be interpreted as the average uncertainty (information) revealed by knowing the value of X given that we know the value of Y.

Lemma 5 (The Chain Rule of Entropy)**.** *For two random variables X and Y:*

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

Furthermore, for the vector of random variables X^n,

$$H(X^n) = H(X_1) + H(X_2|X_1) + H(X_3|X_1,X_2) + \cdots + H(X_n|X_1,X_2,\cdots,X_{n-1}).$$

The chain rule can be interpreted as follows. The amount of information in the pair (X,Y) can be expressed as the amount of information in X plus the amount of information in Y given that X is known.

Proof. We prove the lemma for two random variables:

$$\begin{aligned} H(X,Y) &= -\sum_{x,y\in\mathcal{X}\times\mathcal{Y}} P_{X,Y}(x,y)\log P_{X,Y}(x,y) \\ &= -\sum_{x,y\in\mathcal{X}\times\mathcal{Y}} P_{X,Y}(x,y)\log P_X(x)P_{Y|X}(y|x) \\ &= -\sum_{x,y\in\mathcal{X}\times\mathcal{Y}} P_{X,Y}(x,y)\log P_X(x) - \sum_{x,y\in\mathcal{X}\times\mathcal{Y}} P_{X,Y}(x,y)\log P_{Y|X}(y|x) \\ &= H(X) + H(Y|X). \end{aligned}$$

□

Lemma 6 (Conditioning Reduces Entropy)**.** *For two random variables X and Y:*

$$H(Y|X) \leq H(Y), \text{“} = \text{”} \textit{ iff } X \textit{ and } Y \textit{ are independent.}$$

Proof. We previously showed that $H(X,Y) \leq H(X) + H(Y)$ with equality iff X and Y are independent. The result is proved by noting that $H(X,Y) = H(X) + H(Y|X)$. $\square$