Contents

1. The Khinchin (Shannon?) axioms for entropy

Note all random variables we deal with will be discrete, unless otherwise stated.

1.1. Entropy axioms

Definition 1.1 The **entropy** of a discrete random variable X is a quantity $H(X)$ that takes real values and satisfies Normalisation, Invariance, Extendability, Maximality, Continuity and Additivity.

Axiom 1.2 (Normalisation) If X is uniform on $\{0, 1\}$ (i.e. $X \sim \mathrm{Bern}(1/2)$), then $H(X) = 1$.

Axiom 1.3 (Invariance) If $Y = f(X)$ for some bijection f, then $H(Y) = H(X)$.

Axiom 1.4 (Extendability) If X takes values on a set A, B is disjoint from A, Y takes values in $A \sqcup B$, and for all $a \in A$, $\mathbb{P}(Y = a) = \mathbb{P}(X = a)$, then $H(Y) = H(X)$.

Axiom 1.5 (Maximality) If X takes values in a finite set A and Y is uniformly distributed in A, then $H(X) \leq H(Y)$.

Definition 1.6 The **total variance distance** between X and Y is

$$\sup_E |\mathbb{P}(X \in E) - \mathbb{P}(Y \in E)|.$$

Axiom 1.7 (Continuity) H depends continuously on X (with respect to total variation distance).

Definition 1.8 Let X and Y be random variables. The **conditional entropy** of X given Y is

$$H(X \mid Y) := \sum_y \mathbb{P}(Y = y) H(X \mid Y = y).$$

Axiom 1.9 (Additivity) $H(X, Y) := H((X, Y)) = H(Y) + H(X \mid Y)$.

1.2. Properties of entropy

Lemma 1.10 If X and Y are independent, then $H(X, Y) = H(X) + H(Y)$.

Proof (Hints). Straightforward. □

Proof. $H(X \mid Y) = \sum_y \mathbb{P}(Y = y) H(X \mid Y = y)$ Since X and Y are independent, the distribution of X is unaffected by knowing Y, so $H(X \mid Y = y)$ for all y, which gives the result. (Note we have implicitly used Invariance here). □

Corollary 1.11 If $X_1, ..., X_n$ are independent, then

$$H(X_1, ..., X_n) = H(X_1) + \cdots + H(X_n).$$

Proof (Hints). Straightforward. □

Proof. By Lemma 1.10 and induction. □

Lemma 1.12 (Chain Rule) Let $X_1, ..., X_n$ be RVs. Then

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

Proof (Hints). Straightforward. □

Proof. The case $n = 2$ is Additivity. In general,

$$H(X_1, ..., X_n) = H(X_1, ..., X_{n-1}) + H(X_n \mid X_1, ..., X_{n-1}),$$

so the result follows by induction. □

Lemma 1.13 Let X and Y be RVs. If $Y = f(X)$, then $H(X, Y) = H(X)$. Also, $H(Z \mid X, Y) = H(Z \mid X)$.

Proof (Hints). Consider an appropriate bijection. □

Proof. The map $g : x \mapsto (x, f(x))$ is a bijection, and $(X, Y) = g(X)$, so the first statement follows from Invariance. Also,

$$\begin{aligned} H(Z \mid X, Y) &= H(Z, X, Y) - H(X, Y) \quad \text{by additivity} \\ &= H(Z, X) - H(X) \quad \text{by first part} \\ &= H(Z \mid X) \quad \text{by additivity} \end{aligned}$$

□

Lemma 1.14 If X takes only one value, then $H(X) = 0$.

Proof (Hints). Use that X and X are independent. □

Proof. X and X are independent (verify). So by Lemma 1.10, $H(X, X) = 2H(X)$. But by Invariance, $H(X, X) = H(X)$. So $H(X) = 0$. □

Proposition 1.15 If X is uniformly distributed on a set of size 2^n, then $H(X) = n$.

Proof (Hints). Straightforward. □

Proof. Let $X_1, ..., X_n$ be independent RVs, uniformly distributed on $\{0, 1\}$. By Corollary 1.11 and Normalisation, $H(X_1, ..., X_n) = n$. So the result follows by Invariance. □

Proposition 1.16 If X is uniformly distributed on a set A of size r, then $H(X) = \log_2(r)$.

Proof. By Proposition 1.15, Maximality and Invariance, we have $\lfloor \log_2(r) \rfloor \le H(X)$ (by considering the random variable Y on a set A where Y is uniformly distributed on a subset of A of size $2^{\lfloor \log_2(r) \rfloor}$). Now by Corollary 1.11, we similarly have that $\lfloor k \log_2(r) \rfloor \le H(X_1, ..., X_k) = kH(X)$ for all $k \in \mathbb{N}$, where $X_1, ..., X_k$ are IID and have the same distribution as X. So we have $\frac{1}{k} \lfloor k \log_2(r) \rfloor = H(X)$, and taking the limit as $k \to \infty$ gives $\log_2(r) \le H(X)$.

Following a similar argument, by Proposition 1.15, Maximality, and Invariance, we have $H(X) \le \lceil \log_2(r) \rceil$. By Corollary 1.11, we have that $kH(X) = H(X_1, ..., X_k) \le \lceil k \log_2(r) \rceil$ for all $k \in \mathbb{N}$, which gives $H(X) \le \log_2(r)$. □