Contents

1. The Khinchin axioms for entropy

Note all random variables we deal with will be discrete, unless otherwise stated. We use $\log = \log_2$.

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 the **Khinchin axioms**: Normalisation, Invariance, Extendability, Maximality, Continuity and Additivity.

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

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

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 n, then $H(X) = \log n$.

Proof (Hints). Straightforward. □

Proof. Let $r \in \mathbb{N}$ and let $X_1, ..., X_r$ be independent copies of X. Then $(X_1, ..., X_r)$ is uniform on A^r, and $H(X_1, ..., X_r) = rH(X)$. Now pick k such that $2^k \le n^r \le 2^{k+1}$. Then by Proposition 1.15, Invariance and Maximality, $k \le rH(X) \le k + 1$. So $\frac{k}{r} \le \log n \le \frac{k+1}{r}$ and $\frac{k}{r} \le H(X) \le \frac{k+1}{r}$ for all $r \in \mathbb{N}$. So $H(X) = \log n$, as claimed. □

Theorem 1.17 (Khinchin) If H satisfies the Khinchin axioms and X takes values in a finite set A, then

$$H(X) = \sum_{a \in A} p_a \log(1/p_a) = \mathbb{E}\left[\log \frac{1}{P_X(X)}\right],$$

where $p_a = \mathbb{P}(X = a)$.

Proof (Hints).

- Explain why it is enough to prove for when the p_a are rational.
- Pick $n \in \mathbb{N}$ such that $p_a = \frac{m_a}{n}$, $m_a \in \mathbb{N}_0$. Let Z be uniform on $[n]$. Let $\{E_a : a \in A\}$ be a partition of $[n]$ into sets with $|E_a| = m_a$.

$\square$

Proof. First we do the case where all $p_a \in \mathbb{Q}$. Pick $n \in \mathbb{N}$ such that $p_a = \frac{m_a}{n}$, $m_a \in \mathbb{N}_0$. Let Z be uniform on $[n]$. Let $\{E_a : a \in A\}$ be a partition of $[n]$ into sets with $|E_a| = m_a$. By Invariance, we may assume that $X = a \Leftrightarrow Z \in E_a$. Then

$$\begin{aligned}
\log n = H(Z) = H(Z, X) &= H(X) + H(Z \mid X) \\
&= H(X) + \sum_{a \in A} p_a H(Z \mid X = a) \\
&= H(X) + \sum_{a \in A} p_a \log m_a \\
&= H(X) + \sum_{a \in A} p_a (\log p_a + \log n) \\
&= H(X) + \sum_{a \in A} p_a \log p_a + \log n.
\end{aligned}$$

Hence $H(X) = -\sum_{a \in A} p_a \log p_a$.

The general result follows by Continuity. $\square$

Corollary 1.18 Let X and Y be random variables. Then $0 \le H(X)$ and $0 \le H(X \mid Y)$.

Proof (Hints). Trivial. $\square$

Proof. Immediate consequence of Khinchin. $\square$

Corollary 1.19 If $Y = f(X)$, then $H(Y) \le H(X)$.

Proof (Hints). Straightforward. $\square$

Proof. $H(X) = H(X, Y) = H(Y) + H(X \mid Y)$. But $H(X \mid Y) \ge 0$. $\square$

Proposition 1.20 (Subadditivity) Let X and Y be RVs. Then $H(X, Y) \le H(X) + H(Y)$.

Proof (Hints).

- Let $p_{ab} = \mathbb{P}(X = a, Y = b)$. Explain why it is enough to show for the case when the p_{ab} are rational.
- Pick n such that $p_{ab} = m_{ab}/n$ with each $m_{ab} \in \mathbb{N}_0$. Partition $[n]$ into sets E_{ab} of size m_{ab}. Let Z be uniform on $[n]$.
- Show that if X (or Y) is uniform, then $H(X \mid Y) \le H(X)$ and $H(X, Y) \le H(X) + H(Y)$.

- Let $E_b = \cup_a E_{ab}$ for each b. So $Y = b$ iff $Z = E_b$. Now define an RV W as follows: if $Y = b$, then W is uniformly distributed in E_b. Use conditional independence to conclude the result.

$\square$

Proof. Note that for any two RVs X, Y,

$$\begin{aligned} &H(X,Y) \le H(X) + H(Y) \\ \iff &H(X \mid Y) \le H(X) \\ \iff &H(Y \mid X) \le H(Y) \end{aligned}$$

by Additivity. Next, observe that $H(X \mid Y) \le H(X)$ if X is uniform on a finite set, since $H(X \mid Y) = \sum_y \mathbb{P}(Y = y) H(X \mid Y = y) \le \sum_y \mathbb{P}(Y = y) H(X) = H(X)$ by Maximality. By the above equivalence, we also have $H(X \mid Y) \le H(X)$ if Y is uniform on a finite set. Now let $p_{ab} = \mathbb{P}(X = a, Y = b)$, and assume that all p_{ab} are rational. Pick n such that $p_{ab} = m_{ab}/n$ with each $m_{ab} \in \mathbb{N}_0$. Partition $[n]$ into sets E_{ab} of size m_{ab}. Let Z be uniform on $[n]$. WLOG (by Invariance), $(X, Y) = (a, b)$ iff $Z \in E_{ab}$.

Let $E_b = \cup_a E_{ab}$ for each b. So $Y = b$ iff $Z = E_b$. Now define an RV W as follows: if $Y = b$, then $W \in E_b$, but then W is uniformly distributed in E_b and independent of X (and Z). So W and X are conditionally independent given Y, and W is uniform on $[n]$. Then $H(X \mid Y) = H(X \mid Y, W) = H(X \mid W)$ by conditional independence and by Lemma 1.13 (since W determines Y). Since W is uniform, $H(X \mid W) \le H(X)$.

The general result follows by Continuity. $\square$

Corollary 1.21 $H(X) \ge 0$ for any X.

Proof (Hints). (Without using the formula) straightforward. $\square$

Proof. (Without using the formula). By subadditivity, $H(X \mid X) \le H(X)$. But $H(X \mid X) = 0$. $\square$

Corollary 1.22 Let $X_1, \dots, X_n$ be RVs. Then

$$H(X_1, \dots, X_n) \le H(X_1) + \dots + H(X_n).$$

Proof (Hints). Trivial. $\square$

Proof. Trivial by induction. $\square$

Proposition 1.23 (Submodularity) Let X, Y, Z be RVs. Then

$$H(X \mid Y, Z) \le H(X \mid Z).$$

Proof (Hints). Use that $H(X \mid Y, Z = z) \le H(Z \mid Z = z)$. $\square$

Proof. Either:
1. Use that (Y, Z) determines Z and Corollary 1.19.

2. $H(X \mid Y, Z) = \sum_z \mathbb{P}(Z = z) H(X \mid Y, Z = z) \leq \sum_z \mathbb{P}(Z = z) H(X \mid Z = z) = H(X \mid Z)$.

□

Remark 1.24 Submodularity can be expressed in several equivalent ways. Expanding using Additivity gives

$$H(X, Y, Z) - H(Y, Z) \leq H(X, Z) - H(Z)$$

and

$$H(X, Y, Z) \leq H(X, Z) + H(Y, Z) - H(Z)$$

and

$$H(X, Y, Z) + H(Z) \leq H(X, Z) + H(Y, Z).$$