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

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

Proof (Hints). Straightforward. □

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

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

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 $X_1, \ldots, X_n$ be independent RVs, uniformly distributed on $\{0, 1\}$. By Corollary 1.11 and Normalisation, $H(X_1, \ldots, X_n) = n$. 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 $X_1, \ldots, X_r$ be independent copies of X. Then $(X_1, \ldots, X_r)$ is uniform on A^r, and $H(X_1, \ldots, X_r) = rH(X)$. 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 $0 \leq H(X)$ and $0 \leq H(X \mid Y)$.

Proof (Hints). Trivial. $\square$

Proof. Immediate consequence of Khinchin. $\square$

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

Proof (Hints). Straightforward. $\square$

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

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

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 $H(X \mid Y) \leq H(X)$ and $H(X, Y) \leq H(X) + H(Y)$.
- 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,

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

by Additivity. Next, observe that $H(X \mid Y) \leq H(X)$ if X is uniform on a finite set, since $H(X \mid Y) = \sum_y \mathbb{P}(Y = y) H(X \mid Y = y) \leq \sum_y \mathbb{P}(Y = y) H(X) = H(X)$ by Maximality. By the above equivalence, we also have $H(X \mid Y) \leq H(X)$ 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, $H(X \mid W) \leq H(X)$.

The general result follows by Continuity. $\square$

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

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

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

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

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

Proof (Hints). Trivial. $\square$

Proof. Trivial by induction. $\square$

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

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

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

Proof.

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

$\square$

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

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

Lemma 1.25 Let X, Y, Z be RVs with $Z = f(Y)$. Then $H(X \mid Y) \leq H(X \mid Z)$.

Proof (Hints). Straightforward. □

Proof. We have

$$\begin{aligned} H(X \mid Y) = H(X,Y) - H(Y) &= H(X,Y,Z) - H(Y,Z) \\ &\leq H(X,Z) - H(Z) = H(X \mid Z) \end{aligned}$$

by Submodularity. □

Lemma 1.26 Let X, Y, Z be RVs with $Z = f(X) = g(Y)$. Then

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

Proof (Hints). Straightforward. □

Proof. By Submodularity, we have $H(X,Y,Z) + H(Z) \leq H(X,Z) + H(Y,Z)$, which implies the result, since Z depends on X and Y. □

Lemma 1.27 Let X be an RV taking values in a finite set A and let Y be uniform on A. If $H(X) = H(Y)$, then X is uniform.

Proof (Hints). Use Jensen's inequality. □

Proof. Let $p_a = \mathbb{P}(X = a)$. Then

$$H(X) = \sum_{a \in A} p_a \log(1/p_a) = |A| \cdot \mathbb{E}_{a \in A} p_a \log\left(\frac{1}{p_a}\right).$$

The function $x \mapsto x \log(1/x)$ is concave on $[0,1]$. So by Jensen's inequality,

$$H(X) \leq |A| \cdot (\mathbb{E}_{a \in A} p_a) \cdot \log\left(\frac{1}{\mathbb{E}_{a \in A} p_a}\right) = \log|A| = H(Y),$$

with equality iff $a \mapsto p_a$ is constant, i.e. X is uniform. □

Corollary 1.28 If $H(X,Y) = H(X) + H(Y)$, then X and Y are independent.

Proof (Hints). Go through the proof of Subadditivity and check when equality holds.
□

Proof. We go through the proof of subadditivity and check when equality holds. Suppose that X is uniform on A. Then

$$H(X \mid Y) = \sum_y \mathbb{P}(Y = y) H(X \mid Y = y) \le H(X),$$

with equality iff $H(X \mid Y = y)$ is uniform on A for all y (by Lemma 1.27), which implies that X and Y are independent.

At the last stage of the proof, we said $H(X \mid Y) = H(X \mid Y, W) = H(X \mid W) \le H(X)$, where W was uniform. So equality holds only if X and W are independent, which implies (since Y depends on W), that X and Y are independent. □

Definition 1.29 Let X and Y be RVs. The **mutual information**

$$\begin{aligned} I(X : Y) &:= H(X) + H(Y) - H(X, Y) \\ &= H(X) - H(X \mid Y) \\ &= H(Y) - H(Y \mid X). \end{aligned}$$

Remark 1.30 Subadditivity is equivalent to the statement that $I(X : Y) \ge 0$, and Corollary 1.28 implies that $I(X : Y) = 0$ iff X and Y are independent.

Note that $H(X, Y) = H(X) + H(Y) - I(X : Y)$ (note the similarity to the inclusion-exclusion formula for two sets).

Definition 1.31 Let X, Y, Z be RVs. The **conditional mutual information** of X and Y given Z is

$$\begin{aligned} I(X : Y \mid Z) &:= \sum_z \mathbb{P}(Z = z) I(X \mid Z = z : Y \mid Z = z) \\ &= \sum_z \mathbb{P}(Z = z)(H(X \mid Z = z) + H(Y \mid Z = z) - H(X, Y \mid Z = z)) \\ &= H(X \mid Z) + H(Y \mid Z) - H(X, Y \mid Z) \\ &= H(X, Z) + H(Y, Z) - H(X, Y, Z) - H(Z). \end{aligned}$$

Submodularity is equivalent to the statement that $I(X : Y \mid Z) \ge 0$.

2. A special case of Sidorenko's conjecture

Definition 2.1 Let G be a bipartite graph with (finite) vertex sets X and Y and density α (defined to be $\frac{|E(G)|}{|X| \cdot |Y|}$). Let H be another (think of it as small) bipartite graph with vertex sets U and V and m edges. Now let $\varphi : U \to X$ and $\psi : V \to Y$. We say that (φ, ψ) is a **homomorphism** if $\varphi(x)\varphi(y) \in E(G)$ for every edge $xy \in E(H)$.

Conjecture 2.2 (Sidorenko's Conjecture) For every G, H, for random $\varphi : U \to X$, $\psi : V \to Y$,

$$\mathbb{P}((\varphi, \psi) \text{ is a homomorphism}) \ge \alpha^m.$$

Remark 2.3 Sidorenko's Conjecture is not hard to prove when H is the complete bipartite graph $K_{r,s}$ (the case $K_{2,2}$ can be proved using Cauchy-Schwarz: exercise).

Theorem 2.4 Sidorenko's Conjecture is true if H is a path of length 3.

Proof. We want to show that if G is a bipartite graph of density α with vertex sets X, Y of size m and n, and we choose $x_1, x_2 \in X$, $y_1, y_2 \in Y$ independently at random, then $\mathbb{P}(x_1y_1, y_1x_2, x_2y_2 \in E(G)) \geq \alpha^3$.

It would be enough to let P be a path of length 3 chosen uniformly at random and show that $H(P) \geq \log(\alpha^3 m^2 n^2)$ (by Proposition 1.16). Instead, we shall define a different RV taking values in the set of all paths of length 3 (including degenerate paths). To do this, let (X_1, Y_1) be a random edge of G (with $X_1 \in X$, $Y_1 \in Y$). Now let X_2 be a random neighbour of Y_1 and Y_2 be a random neighbour of X_2. It will be enough to prove that

$$H(X_1, Y_1, X_2, Y_2) \geq \log(\alpha^3 m^2 n^2).$$

We can choose X_1, Y_1 in three equivalent ways:

1. Pick an edge uniformly from all edges
2. Pick a vertex x with probability proportional to its degree $d(x)$, and then a random neighbour Y of x.
3. Same as above with x and y exchanged.

It follows that $Y_1 = y$ with probability $\deg(y)/|E(G)|$, so X_2Y_1 is uniform in $E(G)$, so $X_2 = x'$ with probability $d(x')/|E(G)|$, so X_2Y_2 is uniform in $E(G)$.

Let U_A be the uniform distribution on A. Therefore,

$$\begin{aligned}
H(X_1, Y_1, X_2, Y_2) &= H(X_1) + H(Y_1 \mid X_1) + H(X_2 \mid X_1, Y_1) + H(Y_2 \mid X_1, Y_1, X_2) \\
&= H(X_1) + H(Y_1 \mid X_1) + H(X_2 \mid Y_1) + H(Y_2 \mid X_2) \\
&= H(X_1) + H(X_1, Y_1) - H(X_1) + H(X_2, Y_1) - H(Y_1) + H(X_2, Y_2) - H(Y_2) \\
&= 3H\left(U_{E(G)}\right) - H(Y_1) - H(X_2) \\
&\geq 3H\left(U_{E(G)}\right) - H(U_Y) - H(U_X) \\
&= 3\log(\alpha mn) - \log n - \log m \\
&= \log(\alpha^3 m^2 n^2).
\end{aligned}$$

So we are done, by Maximality. Alternative finish to the proof: let X', Y' be uniform in X, Y and independent of each other and X_1, Y_1, X_2, Y_2. Then

$$\begin{aligned}
H(X_1, Y_1, X_2, Y_2, X', Y') &= H(X_1, Y_1, X_2, Y_2) + H(U_X) + H(U_Y) \\
&\geq 3H\left(U_{E(G)}\right)
\end{aligned}$$

by above. So by Maximality, the number of paths of length 3 times $|X|$ times $|Y|$ is $\geq |E(G)|^3$. □

3. Brigner's theorem

Definition 3.1 Let A be an $n \times n$ matrix over $\mathbb{R}$. The **permanent** of A is

$$\mathrm{per}(A) := \sum_{\sigma \in S_n} \prod_{i=1}^{n} A_{i\sigma(i)},$$

i.e. "the determinant without the signs".

Remark 3.2 Let G be a bipartite graph with vertex sets X, Y of size n. Given $(x, y) \in X \times Y$, let

$$A_{xy} = \begin{cases} 1 & \text{if } xy \in E(G) \\ 0 & \text{if } xy \notin E(G) \end{cases},$$

i.e. A is the bipartite adjacency matrix of G. Then $\mathrm{per}(A)$ is the number of perfect matchings in G.

Brigman's theorem concerns how large $\mathrm{per}(A)$ can be if A is a $0, 1$ matrix and the sum of the entries in the i-th row is d_i.

Example 3.3 (TODO: insert diagram) Let G be a disjoint union of K_{a_i,a_i}'s, $i = 1, ..., k$, with $a_1 + \cdots + a_k = n$. Then the number of perfect matchings in G is $\prod_{i=1}^{k} a_i!$.

Theorem 3.4 (Brigman) Let G be a bipartite graph with vertex sets X, Y of size n. Then the number of perfect matchings in G is at most

$$\prod_{x \in X} (\deg(x)!)^{1/\deg(x)}.$$

Proof (by Radhakrishnan). Each (perfect) matching corresponds to a bijection $\sigma : X \to Y$ such that $x\sigma(x) \in E(G)$ for all $x \in X$. Let σ be chosen uniformly from all such bijections. Then by Chain Rule,

$$H(\sigma) = H(\sigma(x_1)) + H(\sigma(x_2) \mid \sigma(x_1)) + \cdots + H(\sigma(x_n) \mid \sigma(x_1), ..., \sigma(x_{n-1})),$$

where $x_1, ..., x_n$ is some enumeration of X. $H(\sigma(x_1)) \le \log \deg(x_1)$. $H(\sigma(x_2) \mid \sigma(x_1)) \le \mathbb{E}_\sigma \log \deg_{x_1}^\sigma(x_2)$, where $\deg_{x_1}^\sigma(x_2) = |N(x_2) \setminus \{\sigma(x_1)\}|$. In general,

$$H(\sigma(x_i) \mid \sigma(x_1), ..., \sigma(x_{i-1})) \le \mathbb{E}_\sigma \log \deg_{x_1,...,x_{i-1}}^\sigma(x_i),$$

where $\deg_{x_1,...,x_{i-1}}^\sigma(x_i) = |N(x_i) \setminus \{\sigma(x_1), ..., \sigma(x_{i-1})\}|$. □