— Lecture 2 —

- X disc. r.v. over A

entropy $H(X)$ measures amount of
- information
- surprise
- randomness

$$H(X) = \sum_{x \in A} \Pr(X=x) \cdot \log \frac{1}{\Pr(X=x)}$$

Recall Shannon-Khinchin entropy axioms

1. Invariance. $H(X)$ determined by dist of X
2. Maximality. $H(X)$ maximised by uniform dist.
3. Extensibility
4. Additivity $H(X,Y) = H(X) + H(Y|X)$
5. Continuity
6. Normalisation

Basic properties

- independence: X, Y indep. r.v.

$$H(Y|X) = H(Y), \quad H(X,Y) = H(X) + H(Y)$$

- monotonicity (i) $X \sim A$, $Y \sim B \supseteq A$ $\Rightarrow H(X) \leq H(Y)$, "$=$" if $A = B$

(ii) if X determines Y, i.e. $Y = f(X)$ $\Rightarrow H(Y) \leq H(X)$

- non-negativity $\forall X$ r.v., $H(X) \geqslant 0$

Lem 1.7 Let X be a r.v. taking ≥ 2 diff. values w/ positive prob. $\Rightarrow$ $H(X) > 0$.

Pf: • Say X is define over A, and set

$$c = \max_{a \in A} \Pr(X = a) < 1$$

• Fix $\varepsilon > 0$, $\exists$ large n s.t. $c^n < \varepsilon$

Consider i.i.d. copies $X_1, \cdots, X_n$ of X
(indep. identically dist)

max atom prob. of $(X_1, \cdots, X_n) \leq c^n < \varepsilon$
(defined over A^n)

$\Rightarrow$ we can partition $A^n = A_0 \cup A_1$ s.t.

$\Pr(A_i) = \frac{1}{2} \pm \varepsilon$. $(\frac{1}{2} - \varepsilon \leq \cdot \leq \frac{1}{2} + \varepsilon)$

Define Y r.v. s.t. $Y = i$ iff

$$(X_1, \cdots, X_n) \in A_i$$

Continuity $\Rightarrow$ normalisation $H(Y) > 0$

• By indep, $H(X_1, \cdots, X_n) = n H(X)$

• Y is determined by $(X_1, \cdots X_n)$

monotonicity $\Rightarrow$ $0 < H(Y) \leq H(X_1, \cdots, X_n) = n H(X)$ ∎

Lem 1.8 (Chain rule) Let $X_1, \cdots, X_n$ be r.v.

$$\Rightarrow H(X_1, \cdots, X_n) = H(X_1) + H(X_2|X_1) + \cdots + H(X_n|X_1, \cdots, X_{n-1})$$

- Additivity $n=2$, induction.

1.3 Subadditivity & Shearer's lemma.

Lem 1.9 (Dropping conditioning) Let X, Y, Z be r.v.

$$\Rightarrow \quad H(Y|X) \leq H(Y) \quad \cdots\cdots \quad (*)$$

and $H(Z|Y,X) \leq H(Z|Y)$

Pf (of $*$) Consider the special case X unif.

$\Rightarrow \forall b$, $H(X|Y=b) \leq H(X)$ by maximality

$\Rightarrow H(X|Y) \leq H(X)$

- Additivity $\Rightarrow H(Y|X) = H(X,Y) - H(X) \leq H(X,Y) - H(X|Y)$
$= H(Y)$.

- By continuity, may assume X takes rational atom prob. Take $U' \sim [n]$ as in Const. 1.4

($\Pr(X=a) = \frac{m_a}{n}$, $m_a \in \mathbb{N}$)

$$U'|X=a \sim V_a \subseteq [n]$$

$(V_a : a \in A)$ part. of $[n]$ w/ $|V_a| = m_a$

The special case above

$\Rightarrow H(Y|(U'|X=a)) \le H(Y)$

$\Rightarrow H(Y|U',X) \le H(Y)$

- On the other hand, $(U'|X=a) \sim V_a$

$\Rightarrow$ given $X=a$, U' is indep. of Y.

$\overset{\text{indep.}}{\Rightarrow} H(Y|U',X=a) = H(Y|X=a)$

$\Rightarrow H(Y|U',X) = H(Y|X) \le H(Y)$ ▨

- **Lem 1.10** (Subadditivity) Let $X_1, \cdots, X_n$ be r.v.

$\Rightarrow H(X_1, \cdots, X_n) \le \sum_{i \in [n]} H(X_i)$.

Pf. Chain rule

$\Rightarrow H(X_1, \cdots, X_n) = H(X_1) + H(X_2|X_1) + \cdots + H(X_n|X_1, \cdots, X_{n-1})$

$\overset{\text{Dropping cond.}}{\le} H(X_1) + H(X_2) + \cdots + H(X_n)$. ▨

Def Mutual information of X, Y

$I(X;Y) = H(X) + H(Y) - H(X,Y)$

$= H(X) - H(X|Y)$

- X, Y indep. $\Rightarrow I(X;Y) = 0$

Lem (Shearer's lem) Let $\mathcal{F}$ be a family of subsets of $[n]$ w/ possible repeats such that each coordinate $i \in [n]$ is contained by $\geq k$ members of $\mathcal{F}$. Then for a r.v. $(X_1, \cdots, X_n)$

$$H(X_1, \cdots, X_n) \leq \frac{1}{k} \sum_{F \in \mathcal{F}} H(X_F), \text{ where}$$

X_F is the vector $(X_i : i \in F)$

Rmk: Consider $\mathcal{F} = \{\{1\}, \{2\}, \cdots, \{n\}\}$, $k = 1$

$$H(X_1, \cdots, X_n) \leq \sum_{F \in \mathcal{F}} H(X_F) = \sum_{i \in [n]} H(X_i)$$

Lem 1.12 (Shearer's lem, prob. version)

Let F be a random subset of $[n]$ s.t. $\forall i \in [n]$ $\Pr(i \in F) \geq \mu$. Then for a r.v. $(X_1, \cdots, X_n)$

$$H(X_1, \cdots, X_n) \leq \frac{1}{\mu} \mathbb{E}_F H(X_F).$$

Pf:
- Order the random elements in F as $i_1 < \cdots < i_k$
- Chain rule ($i_1 \in [i_2 - 1]$)

$$\Rightarrow H(X_F) = H(X_{i_1}, \cdots, X_{i_k}) = H(X_{i_1}) + H(X_{i_2} | X_{i_1}, \cdots, X_{i_2 - 1}) + \cdots + H(X_{i_k} | X_{i_1}, X_{i_2}, \cdots, X_{i_{k-1}})$$

- Write $H(X_i \mid X_{<i}) = H(X_i \mid X_1, \dots, X_{i-1})$

- Conditioning more only reduces entropy (Lem 1.9)

$\Rightarrow H(X_F) \geq H(X_{i_1}) + H(X_{i_2} \mid X_{<i_2}) + \cdots + H(X_{i_k} \mid X_{<i_k})$

$$\mathbb{E}_F H(X_F) \geq \mathbb{E}_F \sum_{i \in F} H(X_i \mid X_{<i})$$

$$= \sum_{i \in [n]} \Pr(i \in F) \cdot H(X_i \mid X_{<i})$$

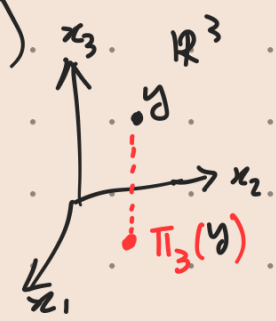
$$\geq \mu \cdot \sum_{i \in [n]} H(X_i \mid X_{<i})$$

$$\overset{\text{Chain rule}}{=} \mu \, H(X_1, \dots, X_n). \quad \blacksquare$$

Application Loomis-Whitney inequality: estimates volume of n-dim body in terms of its $(n-1)$-dim projections.

Def: • a measurable K in $\mathbb{R}^n$, $\mathrm{vol}(K)$ = volume

• $i \in [n]$: $\pi_i : \mathbb{R}^n \to \mathbb{R}^{n-1}$ projection to hyperplane $x_i = 0$: $\pi_i(x_1, \dots, x_n) = (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$



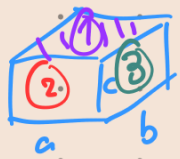
Thm 2.4 (Loomis-Whitney ineq)

Let K be a meas. body in $\mathbb{R}^n$

$$\Rightarrow \mathrm{vol}(K) \le \prod_{i\in[n]} \mathrm{vol}\left(\pi_i(K)\right)^{\frac{1}{n-1}}$$

Rmk Tight for K = axis-aligned box

$n=3$

$K =$ 

Pf: - By standard scaling/limiting arg.

may assume K = union of unit (axis-aligned) boxes

- $\mathrm{vol}(K) = \#$ unit boxes

For each cube $\overset{\text{identify}}{\rightsquigarrow}$ centre

- $X \sim$ boxes in $K \overset{\text{maximality}}{\Rightarrow} H(X) = \log \mathrm{vol}(K)$

Think of $X = (X_1, \cdots, X_n)$, where X_i is the i^{th} coord. of the box X.

$$\log \mathrm{vol}(K) = H(X) = H(X_1, \cdots, X_n) \overset{\text{NTS}}{\le} \frac{1}{n-1} \sum_{i\in[n]} \log \mathrm{vol}\left(\pi_i(K)\right)$$

- Let $\pi \sim \{\pi_1, \pi_2, \cdots, \pi_n\}$ and C be the random set $\subseteq [n]$ recording the set of coordinates π projects to. (e.g. if $\pi = \pi_i$, $C = [n]\setminus\{i\}$)

$Z = (Z_1, \dots, Z_n) \in \mathbb{R}^n$; $\pi(Z) = Z_C := (Z_i : i \in C)$

- As π is unif $\Rightarrow$ $\Pr(i \in C) = \mu = \frac{n-1}{n}$

By Shearer's lem. (Lem 1.12)

$$H(X_1, \dots, X_n) \le \frac{n}{n-1} \mathbb{E}_C H(X_C)$$

$$= \frac{n}{n-1} \mathbb{E}_\pi H(\pi(X))$$

$$= \frac{n}{n-1} \frac{1}{n} \sum_{i=1}^n H(\pi_i(X))$$

$$= \frac{1}{n-1} \sum_{i=1}^n H(\pi_i(X)) \le \frac{1}{n-1} \sum_{i \in [n]} \log \mathrm{vol}(\pi_i(K))$$

Note that $\pi_i(X)$ take values in $\pi_i(K)$

maximality $H(\pi_i(X)) \le \log \mathrm{vol}(\pi_i(K))$

§1.4 Axioms $\iff$ $H(X) = \sum_{a \in A} \Pr(X=a) \log \frac{1}{\Pr(X=a)}$

($\Leftarrow$) Only nontrivial ones $\begin{cases} \text{maximality} \\ \text{additivity} \end{cases}$

Lem 1.13 (Jensen's ineq) X r.v. f concave funct.

$\Rightarrow$ $f(\mathbb{E}X) \ge \mathbb{E}(f(X))$

Prop 1.14 $H(X) = \sum_{x \in A} p_x \cdot \log \frac{1}{p_x}$, $p_x = Pr(X = x)$

Satisfies maximality.

Pf: As $f(x) = \log x$ concave

$$\text{Jensen} \Rightarrow H(X) = \sum_{x \in A} p_x \log \frac{1}{p_x} \leq \log \left(\sum_{x \in A} p_x \frac{1}{p_x} \right) = \log |A|$$

($|A|$ = "")

"=" holds iff all p_x same $\Leftrightarrow X \overset{\text{unif}}{\sim} A$ ▨

Prop 1.15 $H(X) = \sum_{a \in A} p_a \cdot \log \frac{1}{p_a}$, $p_a = Pr(X = a)$

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

Pf: Let Y be defined over B.

$$q_b = Pr(Y = b)$$

$$p_{ab} = Pr(X = a, Y = b)$$

$$q_{b|a} = Pr(Y = b \mid X = a)$$

$$\Rightarrow p_{ab} = Pr(X = a) \cdot Pr(Y = b \mid X = a) = p_a \cdot q_{b|a}$$

$$H(X, Y) = \sum_{a,b} p_{ab} \log \frac{1}{p_{ab}} = \sum_{a,b} p_{ab} \left(\log \frac{1}{p_a} + \log \frac{1}{q_{b|a}} \right)$$

$$1^{st} \text{ term} = \sum_{a,b} p_{ab} \log \frac{1}{p_a} = \sum_{a \in A} \log \frac{1}{p_a} \underbrace{\sum_{b \in B} p_{ab}}_{= p_a} = H(X)$$

Ex: 2^{nd} term $= H(Y|X)$. ▨