2. Fundamentals of Information Theory

Subchapter 2.1 - 2.4 only about discrete distribution.

2.1. Information Measures

Consider 2 random experiment with 3 outcomes, resp., and distributions by

$$(0.9, 0.05, 0.05)$$

$$(0.4, 0.3, 0.3)$$

We seek a measure of

uncertainty (about the outcome) } $\stackrel{!}{=}$
information (gained by the outcome) }

An appropriate measure was introduced by Shannon [48], namely entropy. Formal description & motivation.

The information content of some event E shall only depend on the probability $p = P(E)$.

Measure the information content by a function

$$h : [0,1] \to \mathbb{R}$$

satisfying (i) h continuous on $[0,1]$

(ii) $h(p \cdot q) = h(p) + h(q)$, $p, q \in [0,1]$

(iii) $\exists\, c > 1 : h\left(\frac{1}{c}\right) = 1$

(iii) Normalization, fixing the scale

(ii) Let E_1, E_2 be independent events with prob. p and q, resp., the the event $E_1 \cap E_2$ has uncertainty $h(p) + h(q)$.

If (i), (ii), (iii) hold for some function h, then

$$h(p) = -\log_c(p) \quad , p \in [0,1].$$

Consider description by discrete random variables X (r.v.) with finite support $\mathcal{X} = \{x_1, \dots, x_m\}$ and distribution $P(X = x_i) = p_i$, $p_i \geq 0$, $\sum_{i=1}^{m} p_i = 1$.

Entropy is defined as the average information content of the events $\{X = x_i\}$, $i = 1, \dots, m$.

Def. 2.1.1. Fix $c > 1$, a constant.

$$H(X) = -\sum_{i=1}^{m} P(X = x_i) \log_c P(X = x_i)$$

$$= -\sum_{i=1}^{m} p_i \log_c p_i$$

is called entropy of X (or of $(p_1, \dots, p_m)$). ⊥

Remarks 2.1.2.

a) If $p_i = 0$ for some i, we set $0 \cdot \log 0 = 0$.

b) $H(X)$ only depends on the probabilities, not on the support.

c) The log-base is omitted, however fixed.
Common choices: $c = e = 2.7\ldots$, $c = 2$, $c = 10, \ldots$

d) Let $p(x)$ denote the prob. mass fct. (pmf) or discrete density of X, i.e.,

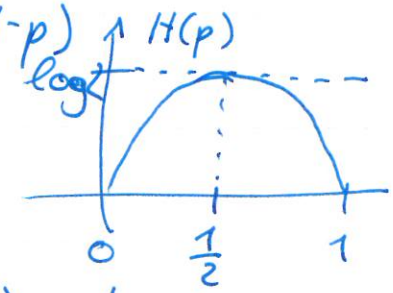
$$p : \mathcal{X} \to [0,1] : x_i \mapsto p(x_i) = p_i$$

Then $H(X) = E\left[\log \frac{1}{p(X)}\right]$.

Examples 2.1.3.

a) $X \sim \mathrm{Bin}(1,p)$, i.e., $P(X=0) = 1-p$, $P(X=1) = p$, $p \in [0,1]$

$H(X) = -p \log p - (1-p)\log(1-p)$



b) $X \sim U(\{1,\ldots,m\})$, i.e., $P(X=k) = \frac{1}{m}$, $k = 1,\ldots,m$.

$$H(X) = -\sum_{i=1}^{m} \frac{1}{m} \log \frac{1}{m} = -\log \frac{1}{m} = \log m$$

Particularly: $m = 26$, $\log_2 26 = 4.7004\ldots$

c) Frequencies of characters in English

A	B	C	...	Z
0.082	0.015	0.028		0.0007

$H(X) = -0.082 \log_2 0.082 - \ldots - 0.0007 \log 0.0007$

$= 4.219$ (< 4.7004 !)

~~Def. 2.1.3~~ Extension to 2-dim discrete random vectors. Let

(X,Y) be a discrete random vector with support $\mathcal{X} \times \mathcal{Y} = \{x_1, \dots, x_n\} \times \{y_1, \dots, y_m\}$ and distribution $P(X=x_i, Y=y_j) = p_{ij}$, $p_{ij} \geq 0$, $\sum_{ij} p_{ij} = 1$.

Def. 2.1.4.

a) $$H(X,Y) = -\sum_{ij} P(X=x_i, Y=y_j) \log P(X=x_i, Y=y_j)$$
$$= -\sum_{ij} p_{ij} \log p_{ij}$$

is called the joint entropy of (X,Y).

b) $$H(X|Y) = -\sum_{j=1}^{d} P(Y=y_j) \sum_{i=1}^{m} P(X=x_i|Y=y_j) \log P(X=x_i|Y=y_j)$$
$$= -\sum_{ij} P(X=x_i, Y=y_j) \log P(X=x_i|Y=y_j)$$

is called <u>conditional entropy</u> of X given Y. (or equivocation).

Th. 2.1.5. (Chain rule)

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

Proof. Denote $p(x_i)$, $p(x_i, y_j)$, $p(y_j|x_i)$ corresponding puf.

$$H(X,Y) = -\sum_{i,j} p(x_i,y_j)\left[\log p(x_i,y_j) - \log p(x_i) + \log p(x_i)\right]$$

$$= -\sum_{i,j} p(x_i,y_j) \log p(y_j|x_i)$$

$$-\sum_i \underbrace{\sum_j p(x_i,y_j)}_{=p(x_i)} \log p(x_i)$$

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

Second equality analogously. ∎

Lemma 2.1.6. (Jensen inequality)

If f is a convex function and X a r.v., then

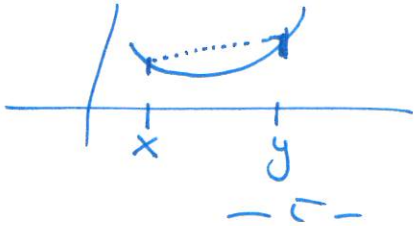
$$E(f(X)) \geq f(EX). \quad \dashv \qquad (*)$$

(*) holds for any r.v. (not only discrete) as long as the expectations are well defined.

---

f is called convex if $f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y)$

$\alpha \in [0,1]$, $x, y \in D(f)$



For discrete r.v. with distribution $(p_1, \dots, p_m)$
(*) reads

$$\sum_{i=1}^{m} p_i f(x_i) \geq f\left(\sum_{i=1}^{m} p_i x_i\right)$$

$$\forall x_1, \dots, x_m \in \mathrm{dom}(f) \quad \forall p_i \geq 0, \sum_{i=1}^{m} p_i = 1.$$

Proof. → Ex.

Lemma 2.1.6. (log-sum inequality)
Let $a_i, b_i \geq 0$, $i = 1, \dots, n$. Then

$$\sum_i a_i \log \frac{a_i}{b_i} \geq \sum_i a_i \log \frac{\sum_j a_j}{\sum_j b_j}$$

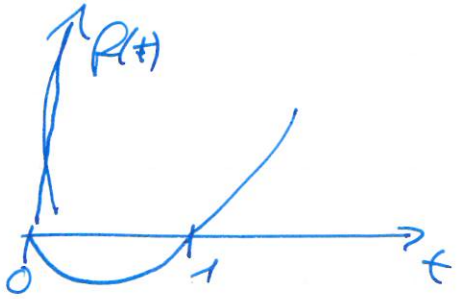
with equality if and only if $\frac{a_i}{b_i} = \text{const.}$ ⊥

We use conventions: $0 \cdot \log 0 = 0$

$a \log \frac{a}{0} = \infty$, if $a > 0$

$0 \log \frac{0}{0} = 0$ (by continuity)

Proof. $f(t) = t \cdot \log t$, $t \geq 0$, is strictly convex.
Since $f''(t) = \frac{1}{t} > 0$, $t > 0$.



w.l.o.g. assume $a_i, b_i > 0$.

By convexity of f

$$\sum_{i=1}^{n} \alpha_i f(t_i) \geq f\left(\sum_{i=1}^{n} \alpha_i t_i\right), \quad \alpha_i \geq 0, \ \sum_{i=1}^{n} \alpha_i = 1$$

Setting $\alpha_i = \frac{b_i}{\sum_{j=1}^{n} b_j}$, $t_i = \frac{a_i}{b_i}$ it follows

$$\sum_i \frac{b_i}{\sum_j b_j} \frac{a_i}{b_i} \log \frac{a_i}{b_i} \geq \sum_i \frac{b_i}{\sum_j b_j} \frac{a_i}{b_i} \log\left(\sum_i \frac{b_i}{\sum_j b_j} \frac{a_i}{b_i}\right)$$

$$\Leftrightarrow \quad \frac{\sum_i a_i \log \frac{a_i}{b_i}}{\sum_j b_j} \geq \frac{\sum_i a_i}{\sum_j b_j} \log \frac{\sum_i a_i}{\sum_j b_j}$$ ∎

Corollary 2.1.7. Let $p = (p_1, \dots, p_n)$, $q = (q_1, \dots, q_n)$ be stoch. vectors, i.e., $p_i \geq 0$, $q_i \geq 0$, $\sum_i p_i = \sum_i q_i = 1$. Then

$$-\sum_{i=1}^{n} p_i \log p_i \leq -\sum_{i=1}^{n} p_i \log q_i$$

"=" if and only if $p = q$. ∎

Proof. In Lo. 2.1.6. set

$a_i = p_i$, $b_i = q_i$ and observe that $\sum_i p_i = \sum_i q_i = 1$ ∎