INFORMATION THEORY

Literature:
- Cover & Thomas „Elements of info. theory."
- Mackay „Info. Theory, Inference & Learning Algorithms" (free online copy)
- Shannon „The Math. Theory of Communication" (1949)

some history:

- belief ~'40: sending info. at positive rates is not possible with negligible error.

- Shannon '48: - arbitrary small error probability is achievable for all rates below "capacity". The latter can be computed and is essentially always non-zero.
  - signals have irreducible complexity below which they cannot be compressed.

  (crucial idea in both cases: description of signal/info. as random processes)

  '49 - Shannon-Nyquist sampling theorem
  - foundations of modern cryptography

modern applications:
- data compression
  - lossless (ZIP, gzip, Dolby True HD, ...)
  - lossy (JPEG, MP3, ...)
- error correction: CD, DVD, Blue-ray, bar codes, ...
- channel coding: satellite communication, WLAN, mobile networks, ...

future applications:
- quantum information theory?

I. Preliminaries

I.1. Probability theory

- $\mathcal{X}$ finite set (symbols, events, ...)
- X random variable with range in $\mathcal{X}$ and distribution

  $p: \mathcal{X} \to \mathbb{R}_+ := [0,\infty)$

  $\sum_{x\in\mathcal{X}} p(x) = 1$ (we also use $p(x) = p_X(x) = p_x$)
- expectation value $E(X) := \sum_{x\in\mathcal{X}} x\, p(x)$

  (if $\mathcal{X}$ is embedded in a linear space)
- joint distribution $p_{XY}: \mathcal{X}\times\mathcal{Y} \to \mathbb{R}_+$ for vector-valued r.v.s
  - marginals $p_X(x) := \sum_{x\in\mathcal{Y}} p(x,y)$, $p_Y(y) := \sum_{x\in\mathcal{X}} p(x,y)$
  - conditional distribution $p(x|y) := \frac{p(x,y)}{p_Y(y)}$ for $p_Y(y) \neq 0$
  - X & Y are independent r.v.s iff $p(x,y) = p_X(x)\,p_Y(y)\ \forall x,y$

    $\Leftrightarrow p(x|y) = p_X(x)\ \forall x\ \forall y: p_Y(y) > 0$

I.2. Convexity

Def.: • Let V be an $\mathbb{R}$-vector space. $C \subseteq V$ is a "convex set" iff

$$\forall\, \lambda\in[0,1]: \left(x,y\in C \Rightarrow \lambda x + (1-\lambda)y \in C\right)$$

- Let C be a convex set. $f: C \to \mathbb{R}$ is a "convex function" on C

iff $\forall x,y\in C\ \forall \lambda\in[0,1]$:

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$$

- f is called "strictly convex" iff '=' holds only if $\lambda\in\{0,1\}$ or $x=y$.
- f is "(strictly) concave" iff $-f$ is (strictly) convex

Lemma: Let $C \subseteq \mathbb{R}^n$ be convex and open and $f \in \mathcal{C}^2(C, \mathbb{R})$. Then

(i) $\forall x \in C: f''(x) \geq 0 \iff f$ convex on C

(ii) $\forall x \in C: f''(x) > 0 \Rightarrow f$ strictly convex on C

Lemma: (Jensen's inequality) If X is a real valued r.v. and $f: \mathbb{R} \to \mathbb{R}$ convex, then $\mathbb{E}(f(X)) \geq f(\mathbb{E}(X))$.

proof: by induction on $n = |\mathcal{X}|$ with $n=2$ the definition of convexity ... □

II. Entropic quantities

II.1. Entropy as measure of uncertainty

"Bar Kochba game": identify $x \in \mathcal{X}$ with minimal number n of binary questions.

- necessary: $2^n \geq |\mathcal{X}_0|$, $\mathcal{X}_0 := \{x \in \mathcal{X} \mid p(x) > 0\}$
- $\lceil x \rceil := \min\{n \in \mathbb{Z} \mid n \geq x\}$
- $\log x := \log_2 x$
- $\lceil \log |\mathcal{X}_0| \rceil$ questions are sufficient by partitioning $\mathcal{X}_0$ according to binary tree.
- take m independent copies $\mathcal{X}_0^m$. On average (i.e., per copy)

$$-\frac{1}{m} + \log|\mathcal{X}_0| \leq \frac{\lfloor \log |\mathcal{X}_0|^m \rfloor}{m} \quad n \leq \frac{\lceil \log |\mathcal{X}_0|^m \rceil}{m} \leq \frac{1}{m} + \log|\mathcal{X}_0|$$

$$\xrightarrow{m \to \infty} \log|\mathcal{X}_0| =: H_0(X) \quad \text{"Hartley entropy"}$$

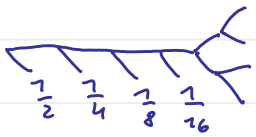
or "0-Renyi entropy"

However, this does not take prob.s of the events into account.

Exp.: $p = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right)$

$\Rightarrow H_0(X) = 3$

But on average 2 questions are sufficient if we act according to



→ average # of questions: $\frac{1}{2} + 2\frac{1}{4} + 3\frac{1}{8} + 4\frac{1}{16} + 4 \cdot 6 \cdot \frac{1}{64} = 2$

Shannon entropy $H(X) := -\sum_{x;\, p(x)>0} p(x) \log p(x)$

$= -\sum_{x} p(x) \log p(x)$ with $0 \log 0 := 0$

consider H as functional on $\mathcal{P} := \bigcup_{n \in \mathbb{N}} \left\{ p \in \mathbb{R}_+^n \mid \sum_x p_x = 1 \right\}$

properties:

(i) symmetry: $H(p_1, \dots, p_n) = H(p_{\pi(1)}, \dots, p_{\pi(n)}) \quad \forall \pi \in S_n$

(ii) expansibility: $H(p_1, \dots, p_n, 0) = H(p_1, \dots, p_n)$

(iii) additivity: $H(XY) = H(X) + H(Y)$ if X, Y independent

(iv) subadditivity: $H(XY) \leq H(X) + H(Y)$

proof:

(iii) $-\sum_{x,y} p_x q_y \log(p_x q_y) = -\sum_{x,y} p_x q_y \left(\log p_x + \log q_y\right)$

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

(iv) $H(X) + H(Y) - H(XY) = \sum_{x,y} p(x,y)\left[\log p(x,y) - \log p(x) - \log p(y)\right]$

$= -\sum_{x,y} p(x,y) \log\left[\frac{p(x)p(y)}{p(x,y)}\right]$

- Jensen
- $-\log$ convex

$\geq -\log \sum_{x,y} p(x)p(y) = 0$

□

Thm.: (axiomatic characterization of entropies)

Let $h: \underline{P} \to \mathbb{R}$ be a functional satisfying (i)–(iv),

then $\exists\, a, b \in \mathbb{R}_+$: $\quad h = a H_0 + b H$

If in addition $h\left(\frac{1}{2}, \frac{1}{2}\right) = 1$ and $\lim_{p \searrow 0} h(p, 1-p) = 0$, then $h = H$.

proof: [Aczél, Forte, Ng 1974]