Information and complexity

Probabilistic and algorithmic foundational aspects

Dimitri Petritis

Institut de recherche mathématique de Rennes
Université de Rennes 1 et CNRS (UMR 6625)

September 2022

Information and complexity

Probabilistic and algorithmic foundational aspects

Dimitri Petritis

Institut de recherche mathématique de Rennes
Université de Rennes 1 et CNRS (UMR 6625)

September 2022

Information vs. complexity

What is "information" and how is related to algorithmic problems

$\{0,1\}$
$\{a, \ldots z\}$
$\{A, C, G, T\}$

Statement of the problem: X random variable with outcomes in finite alphabet $\mathbb{X}$.

- $\mathbf{X} = X_1 \cdots X_n$ random sequence = message = n-letter random word on $\mathbb{X}$.
- Ignore semantics of message.
- Concentrate on how to efficiently exchange message between sender and receiver.

Information content = decrease of uncertainty when the outcome of the r.v. is revealed [Shannon1948].

- Information = probabilistic notion; concerns ensembles of messages.
- Closely related to thermodynamic entropy [Boltzmann1896], introduced to explain macroscopic irreversibility despite microscopic reversibility in theory of gases.

Complexity = length of the shortest programme from which sequence can be reproduced. Notion independently introduced in [Solomonoff1964, Kolmogorov1965, Chaitin1974].

- Complexity = algorithmic notion; concerns individual messages.
- A.k.a. algorithmic information.
- For ergodic processes: complexity per symbol = entropy per symbol. [Horibe2003].

Plan of the course

1. Basic postulates, definition, significance, and properties of information content; related functions.
2. Source coding; compression algorithms.
3. Channel coding; fundamental theorem of information transmission.
4. Kolmogorov complexity; Turing machine description of informational content.

Information

$x \in \mathbb{X}$

- X random variable taking values in finite alphabet $\mathbb{X}$.
- Law of X: $\mathbb{P}_X(\{x\}) = \mathbb{P}(X = x) = \mathbf{p}(x)$, determined by $\mathbf{p} = (p(x))_{x \in \mathbb{X}}$, its probability vector.

$p(x) \geq 0$

$\sum_{x \in \mathbb{X}} p(x) = 1$

Function $h =$ quantifier of decrease of uncertainty of event $\{X = x\}$ $(x \in \mathbb{X}.)$

$\mathbb{X} = \{0,1\}$
$p(0) = p(1) = 1/2$

Uncert before − uncert after = 0

- Before r.v. X has been observed: maximal (a priori) uncertainty.
- After observation of the outcome: null uncertainty.
- Hence: reduction of uncertainty = a priori uncertainty.
- Intuitively: uncertainty of $\{X = x\}$ = function of $p(x)$.
- Define for fixed event $A \subset \mathbb{X}$ — having $\mathbb{P}_X(A) = \cancel{p}\ q$ — the function h by
$$]0,1] \ni \cancel{p}\ q \mapsto h(q) \in \mathbb{R}.$$

Function $H =$ expectation of h. Since h varies a lot, compute expectation (average uncertainty):

$$H(\mathbf{p}) = \sum_{x \in \mathbb{X}} p(x) h(p(x)).$$

UNIVERSITÉ DE RENNES 1

Information
Basic postulates

$$H_M(p) = \sum_{x=1}^{n} p(x)\, h(p(x))$$

- Fix henceforth $M := |\mathbb{X}|$ and write **provisionally** $H_M(\mathbf{p})$ instead of $H(\mathbf{p})$.
- Easier to (correctly) guess outcome of coin tossing than number lotery.

Postulate (of monotonicity)

Function $f : \mathbb{N} \to \mathbb{R}_+$ defined by $f(M) := H_M((\frac{1}{M}, \ldots, \frac{1}{M}))$ is strictly increasing.

$\mathbb{X} = \{1, \ldots M\}$

$\underline{p} = (\frac{1}{M}, \ldots \frac{1}{n})$

Information

Basic postulates (cont'd)

- Let X and Y independent r.v. uniformly distributed in $\mathbb{X}$ and $\mathbb{Y}$, with $|\mathbb{X}| = L$ and $|\mathbb{Y}| = M$.
- Composite experiment described by r.v. $(X, Y) \in \mathbb{X} \times \mathbb{Y}$, with $|\mathbb{X} \times \mathbb{Y}| = L \times M$.
- If outcome of X revealed, uncertainty of Y unaffected. However, total uncertainty $f(LM)$ decreased by $f(L)$.
- Hence

Postulate (of extensivity)

For all $L, M \geq 1$, $f(LM) = f(L) + f(M)$.

Information
Basic postulates (cont'd)

- Relaxing uniformity of distribution: **p arbitrary** probability vector on $\mathbb{X}$ ($|\mathbb{X}| = M$)
$$\mathbb{P}(X = x) = p(x), x \in \mathbb{X}.$$
- Partition $\mathbb{X} = \mathbb{X}_1 \sqcup \mathbb{X}_2$; let $q_i = \sum_{x \in \mathbb{X}_i} p(x)$, with $|\mathbb{X}_i| = M_i, i = 1, 2.$
- Split random experiment into two steps:

$$\begin{aligned}\mathbb{P}(X = x) &= \mathbb{P}(X = x | X \in \mathbb{X}_1)\mathbb{P}(X \in \mathbb{X}_1) + \mathbb{P}(X = x | X \in \mathbb{X}_2)\mathbb{P}(X \in \mathbb{X}_2)\\ &= \frac{p(x)}{q_1} q_1 \mathbb{1}_{\mathbb{X}_1}(x) + \frac{p(x)}{q_2} q_2 \mathbb{1}_{\mathbb{X}_2}(x)\\ &= p(x).\end{aligned}$$

Postulate (of grouping)

With above notation,
$$H_M(\mathbf{p}) = q_1 H_{M_1}\left(\frac{\mathbf{p}\restriction_{\mathbb{X}_1}}{q_1}\right) + q_2 H_{M_2}\left(\frac{\mathbf{p}\restriction_{\mathbb{X}_2}}{q_2}\right) + H_2((q_1, q_2)).$$

$p(x) \geq 0$
$\sum p(x) = 1$

$P(X = x | X \in \mathbb{X}_1) = \frac{P(X = x, X \in \mathbb{X}_1)}{P(\mathbb{X}_1)}$

$\mathbb{X} = \{1, 2, 3, 4, 5\}$
$p = (p_1, p_2, p_3, p_4, p_5)$
$\mathbb{X}_1 = \{1, 2\}$
$\mathbb{X}_2 = \{3, 4, 5\}$
$p\restriction_{\mathbb{X}_1} = (p_1, p_2)$
$p\restriction_{\mathbb{X}_2} = (p_3, p_4, p_5)$

Information
Basic postulates (cont'd)

- Smooth dependence of H_2 on $\mathbf{p}$.

Postulate (of continuity)

Function $H_2(p, 1-p)$ is continuous in $p \in [0,1]$.

Information

Existence and uniqueness of function verifying previous postulates

Theorem

The unique function verifying the four previous postulates is the function

$$PV_M \ni \mathbf{p} \mapsto H_M(\mathbf{p}) = -C \sum_{x \in \mathbb{X}} p(x) \log p(x),$$

where the logarithm is in arbitrary base $b > 1$, and $PV_M =$ the set of probability vectors of dimension M.

$PV_M = \{ \mathbf{p} = (p_1 \cdots p_M) \in \mathbb{R}_+^M$
$\sum p_i = 1 \}$

1-A: Proof of the theorem

Remark

H depends on the probability vector (law of X), not X. Nevertheless, write often $H(X)$ to denote $H(\mathbf{p})$, where $\mathbf{p}$ law of X.

Prove: $H(\underline{p}) = -c \sum_{x \in \mathcal{X}} p_x \log p_x$

Information $\log_2$

Physics $\ln$

$a, b \geq 1$

$x > 0$

$\log_a x = (\log_a b) \log_b x$ —

$a = 10 \quad b = e$

$\log_{10} x = y \Rightarrow x = 10^y \quad \log_e x = \log_e 10^y$

?

$= \log_e 10^{\log_{10} x}$

Step 1: For integers $M, k \geq 1$ by extensivity postulate

$f(M^k) = f(M \cdot M^{k-1}) = f(M) + f(M^{k-1})$

iterate: $f(M^k) = k \cdot f(M)$

Step 2: Shall show: for integer $M \geq 1$, $\exists C > 0$ s.t. ;
there exists

$f(M) = C \log M$ (*)

$M = 1$: $f(1) = f(1 \cdot 1) \overset{\text{extensivity}}{=} f(1) + f(1) \Rightarrow f(1) = 0$ * ok for $M = 1$

~~Suppose that (*) true up to $M>1$~~: $f(M) = C \log M$.

$\forall$ integer $r \geq 1$, $\exists k \geq 0$ integer s.t. $M^k \leq 2^r < M^{k+1}$ (one will be strict)

monotonicity: $f(M^k) \leq f(2^r) \leq f(M^{k+1})$

$$kf(M) \leq rf(2) \leq (k+1)f(M)$$

$$\frac{k}{r} \leq \frac{f(2)}{f(M)} \leq \frac{k+1}{r} \qquad (**)$$

On the other hand: log in a base $b>1$ is strictly increasing

(***) dividing by $\log M$

$$\frac{k}{r} \leq \frac{\log 2}{\log M} \leq \frac{k+1}{r}$$

$$M^k \leq 2^r \leq M^{k+1} \Rightarrow k \log M \leq r \log r \leq (k+1) \log M$$

(**)

$$\left| \frac{\log 2}{\log M} - \frac{f(2)}{f(M)} \right| \leq \frac{1}{r}$$

Since M fixed, r arbitrary $\Rightarrow \frac{f(2)}{f(M)} = \frac{\log 2}{\log M} \Rightarrow f(M) = c \log M$ //

can take $\lim_{r \to \infty}$

$c = \frac{f(2)}{\log 2}$

Since $f(1) = 0$, $f \uparrow$ strictly $\Rightarrow f(2) > 0 \Rightarrow c > 0$

$\underline{\text{Step 3}}$: for $p \in \mathbb{Q} \cap [0,1]$ shall show

$$H_2((p, 1-p)) = -C[p\log p + (1-p)\log(1-p)]$$

$p = \frac{r}{s}$, r, s integers ≥ 1

$$f(s) = H_s\Big(\big(\underbrace{\tfrac{1}{s}, \dots \tfrac{1}{s}}_{r}, \underbrace{\tfrac{1}{s}, \dots \tfrac{1}{s}}_{s-r}\big)\Big)$$

$$= H_2\left(\frac{r}{s}, \frac{s-r}{s}\right) + \frac{r}{s} f(r) + \frac{s-r}{s} f(s-r) \quad \text{by grouping}$$

$$\overset{\text{Step 2}}{=} H_2((p, 1-p)) + C p \log r + C(1-p)\log(s-r)$$

$$= C \log s$$

$$\Rightarrow H_2(p, 1-p) = C\left[p \log r + (1-p)\log(sr) - (p + 1-p)\log s \right]$$

$$= C\left[-p\log p - (1-p)\log(1-p) \right]$$

<u>Step 4</u> : $p \in \,]0,1[$ arbitrary

p can be approximated by seq. of rationals (p_n)

$$p_n \to p$$

$$H_2(p_n, 1-p_n) = -C\left[p_n \log p_n + (1-p_n)\log(1-p_n)\right]$$

$$\downarrow$$

$$H_2(p, 1-p) = -C\left(p\log p + (1-p)\log(1-p)\right)$$

Step 3 : $|\mathcal{X}| = M$

$\underline{p} = (p_1 \dots p_M)$

Must show $H_M(\underline{p}) = -C \sum_{x \in \mathcal{X}} p_x \log p_x$

Formula supposed correct up to $M-1$. $\quad q = p_1 + \dots + p_{M-1}$

$$H_M(\underline{p}) = H_2(q, p_M) + q H_{M-1}\left(\left(\frac{p_1}{q}, \dots \frac{p_{M-1}}{q}\right)\right) + p_M H_1((1))$$

$$= -C\Big[q \log q + p_M \log p_M + q \sum_{k=1}^{M-1} \frac{p_k}{q} \log \frac{p_k}{q} + 0$$

$$+ \sum_{k=1}^{M-1} p_k \log p_k - \Big(\sum_{k=1}^{M-1} p_k\Big) \log q$$

$$\boxed{H_M(\underline{p}) = -C \sum_{i=1}^{M} p_i \log p_i}$$

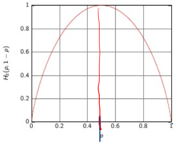
Information
Definition

Definition

The function verifying the basic postulates (with convention $0 \log 0 = 0$)

$$H(\mathbf{p}) := H_{\dim(\mathbf{p})}(\mathbf{p}) = -\sum_{i=1}^{\dim(\mathbf{p})} p_i \log p_i,$$

whose existence and uniqueness established in previous theorem, is the **entropy** or (quantity of) **information** associated with $\mathbf{p} \in \mathrm{PV}$.



$\underline{p} = (p, 1-p)$

Figure: Function H plays important rôle. Behaviour of $H_2(p, 1-p)$ as function of $p \in [0, 1]$.

Historical remarks

- In almost all computer science books, definition of H usually attributed to Claude Elwood Shannon[1].
- Effectively, Shannon's article (1948), establishes — for the first time rigorously — existence, uniqueness, and mathematical properties of information.
- But formula $H = -\sum_i p_i \log p_i$, established 3/4 of a century earlier, in 1877, by Ludwig Eduard Boltzmann[2].
- Next slide: facsimilé of page 41 of Boltzmann's book *Vorlesungen über Gastheorie* (published in 1896.)

[1] Michigan 1916 – Massachussets 2001, American electrical engineer and mathematician, founder of information theory.

[2] Vienna 1844 – Trieste 1906, Austrian physicist, founder of statistical mechanics and defender of atomic theory.

$$Z = \frac{\ldots}{\left(\frac{n}{2}\right)!\left(\frac{n}{2}\right)!} \text{ u. s. w.}$$

Da nun die Anzahl der Moleküle eine überaus grosse ist, so sind $n_1\,\omega$, $n_2\,\omega$ u. s. w. ebenfalls als sehr grosse Zahlen zu betrachten.

Wir wollen die Annäherungsformel:

$$p! = \sqrt{2\,p\,\pi}\left(\frac{p}{e}\right)^p$$

benützen, wobei e die Basis der natürlichen Logarithmen und p eine beliebige grosse Zahl ist.[1])

Bezeichnen wir daher wieder mit l den natürlichen Logarithmus, so folgt:

$$l[(n_1\,\omega)!] = (n_1\,\omega + \tfrac{1}{2})\,l\,n_1 + n_1\,\omega\,(l\,\omega - 1) + \tfrac{1}{2}(l\,\omega + l\,2\,\pi).$$

Vernachlässigt man hier $\frac{1}{2}$ gegen die sehr grosse Zahl $n_1\,\omega$ und bildet den analogen Ausdruck für $(n_2\,\omega)!$, $(n_3\,\omega)!$ u. s. f., so ergibt sich:

$$\boxed{l\,Z = -\,\omega\,(n_1\,l\,n_1 + n_2\,l\,n_2\ldots) + C,}$$

wobei

$$C = l(n!) - n\,(l\,\omega - 1) - \frac{\zeta}{2}(l\,\omega + l\,2\,\pi)$$

für alle Geschwindigkeitsvertheilungen denselben Werth hat, also als Constante zu betrachten ist. Denn wir fragen ja bloss nach der relativen Wahrscheinlichkeit der Eintheilung der verschiedenen Geschwindigkeitspunkte unserer Moleküle in unsere Zellen ω, wobei selbstverständlich die Zelleneintheilung, daher auch die Grösse einer Zelle ω, die Anzahl der Zellen ζ und die Gesammtzahl n der Moleküle und deren gesammte lebendige Kraft als unveränderlich gegeben betrachtet werden müssen. Die wahrscheinlichste Eintheilung der Geschwindig-

(1) Entropy is ...

... an expectation that makes us older

$H(p) = -C \sum p_x \log p_x$ information

$\mathbb{P}(X=x) = p_x$

$(p_1 \cdots p_n)$

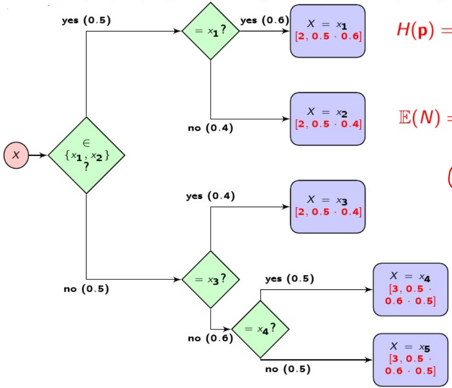
- X an $\mathbb{X}$-valued random variable, whose law $\mathbb{P}_X$ determined by probability vector $\mathbf{p} \in \mathrm{PV}_M$.
- Define random variable $\xi = -\log p(X)$. Then

$$\mathbb{E}\xi = \mathbb{E}\log p(X) = -\sum_{x \in \mathbb{X}} \mathbb{P}(X = x) \log p(x) = -\sum_{x \in \mathbb{X}} p(x) \log p(x) = H(\mathbf{p}).$$

- Boltzmann related irreversibility with positive entropy rate.

(2) Entropy is . . .
. . . the mean number of questions needed to determine X

$X \in \mathbb{X} = \{x_1, \dots, x_5\}$ with $\mathbf{p} = (0.3, 0.2, 0.2, 0.15, 0.15)$.



$H(\mathbf{p}) = 2.27$

$\mathbb{E}(N) = 2.3$

$(X_1, X_2) \in \mathbb{X} \times \mathbb{X}$

$\mathbb{P}(X_1 = x, X_2 = y) =$

$p(x)\, p(y) = \kappa(x,y)$

$H(\kappa) = 2 \times 2.27$

$\mathbb{E}N \in [2 \times 2.27, 2 \times 2.3]$

UNIVERSITÉ DE RENNES 1

(3) Entropy is ...
... the logarithmic ratio of typical over total configurations

$X \in \mathbb{A}$ $\mathbb{A} = \{0,1\}$.

$p(0) = \mathbb{P}(X=0)$
$p(1) = \mathbb{P}(X=1)$

- X random variable in finite alphabet[3] $\mathbb{A}$.
- Law of X given by probability vector $\mathbf{p} := (p_a)_{a \in \mathbb{A}}$.
- For $n \in \mathbb{N}$, consider random variable $\mathbf{X} := \mathbf{X}^{(n)} := (X_1, \ldots, X_n)$, n independent copies of X. Obviously, $\mathbf{X}^{(n)} \in \mathbb{A}^n$, i.e. $\mathbf{X}^{(n)}$ is a random n-letter word over alphabet $\mathbb{A}$. $\in \mathbb{A}^n$
- For fixed n-letter word $\boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_n) \in \mathbb{A}^n$ and letter $a \in \mathbb{A}$, denote by

$$\nu_a(\boldsymbol{\alpha}) := \nu_a^{(n)}(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \mathbb{1}_{\{a\}}(\alpha_k) \in \{0, \ldots, n\}.$$

- Due to independence of $(X_i)_{i=1,\ldots,n}$,

$$\mathbb{P}\left(\mathbf{X}^{(n)} = \boldsymbol{\alpha}\right) = \prod_{a \in \mathbb{A}} p_a^{\nu_a(\boldsymbol{\alpha})}, \quad \forall \boldsymbol{\alpha} \in \mathbb{A}^n$$

$$\mathbb{P}\left(\nu_a^{(n)}(\mathbf{X}^{(n)}) = \ell\right) = C_n^\ell p_a^\ell (1-p_a)^{n-\ell}, \quad \forall a \in \mathbb{A}, \quad \ell = 0, \ldots, n.$$

1-B: Example.

UNIVERSITÉ DE RENNES 1

[3] If $M = \mathrm{card}\mathbb{A}$, w.l.o.g., can suppose $\mathbb{A} \simeq \{1, \ldots, M\}$ and $\mathbf{p} = (p_i)_{i=1,\ldots,M}$.

$n = 5 \qquad \mathbb{A} = \{0,1\} \qquad \underline{\alpha} = 10011$

$a = 0$

$$\left(\frac{\nu_0(\alpha)}{n}, \dots, \frac{\nu_{|\mathbb{A}|}(\alpha)}{n} \right)$$

$$\nu_0(\underline{\alpha}) = \sum_{k=1}^{5} \mathbb{1}_{\{0\}}(\alpha_k) = 2.$$

$$\nu_1(\alpha) = \sum_{k=1}^{5} \mathbb{1}_{\{1\}}(\alpha_k) = 3.$$

$$\boxed{\sum_{a \in \mathbb{A}} \nu_a(\underline{\alpha}) = n}$$

$$\frac{1}{n} \sum_a \nu_a(\alpha) = 1$$

$\underline{X} = (X_1 \dots X_5)$ independent identically distributed (i.i.d) r.v st. $\mathbb{P}(X_k = a) = p_a$

$$\mathbb{P}(\underline{X} = \underline{\alpha}) = \mathbb{P}((X_1 \dots X_5) = \alpha_1 \dots \alpha_5) = \mathbb{P}(X_1 = \alpha_1, \dots, X_5 = \alpha_5)$$

$$= \mathbb{P}(X_1 = \alpha_1) \cdots \mathbb{P}(X_5 = \alpha_5)$$

$$= p_1 p_0 p_0 p_1 p_1 = p_0^{\nu_0(\alpha)} \, p_1^{\nu_1(\alpha)}$$

$\mathbb{P}(\nu_a(\underline{X}^{(n)}) = \ell)$

$\ell \in \{0, \dots, n\}$

$X_1 \cdots X_n$

$\underline{X} =$ ABRACADABRA

$\nu_Z(\underline{X}) = 0$

$\nu_Y(\underline{X}) = 0$

$\nu_A(\underline{X}) = 5$

$A = \{0, 1\}$

$A^n_{1,\ell} = \left\{ \underline{\alpha} = \alpha_1 \dots \alpha_n, \sum_{i=1}^{n} \alpha_i = \ell \right\}$

$|A^n_{1,\ell}| = C^\ell_n = \binom{n}{\ell}$

$$\boxed{\mathbb{P}(\nu_a(\underline{X}^n) = \ell) = |A^n_{1,\ell}| \, p_1^\ell (1-p_1)^{n-\ell}}$$

(3) Entropy is . . .

. . . the logarithmic ratio of typical over total configurations (cont'd)

Remark

More convenient to work with infinite words $\boldsymbol{\alpha} \in \mathbb{A}^{\mathbb{N}}$ or infinite random sequences $\mathbf{X} = (X_1, X_2, \ldots)$ and define

$$\nu_a^{(n)}(\boldsymbol{\alpha}) = \sum_{k=1}^{n} \mathbb{1}_{\{a\}}(\alpha_k).$$

Use notation $\boldsymbol{\alpha}\restriction_n$ or $\mathbf{X}\restriction_n$ to denote restriction of sequence to n first letters. For every $n \in \mathbb{N}_>$ and $\boldsymbol{\alpha} \in \mathbb{A}^{\mathbb{N}}$:

$$\frac{1}{n}\boldsymbol{\nu}^{(n)}(\boldsymbol{\alpha}) \in \mathrm{PV}_{\mathbb{A}}, \quad \text{where} \quad \boldsymbol{\nu}^{(n)}(\boldsymbol{\alpha}) = (\nu_a^{(n)}(\boldsymbol{\alpha}))_{a\in\mathbb{A}}.$$

This probability vector called **type** of $\boldsymbol{\alpha}$.

Typical configurations

Definition

Let $n \geq 1$ integer, $\mathbb{A}$ finite alphabet, $\mathbf{p} \in \mathrm{PV}_{\mathrm{card}\mathbb{A}}$ probability vector, and $K > 0$ integer. An n-letter word $\boldsymbol{\alpha} \in \mathbb{A}^n$ is called **typical** (more precisely $(n, \mathbf{p}, K)$-typical) if

$$\forall a \in \mathbb{A}, \left| \frac{\nu_a^{(n)}(\boldsymbol{\alpha}) - np_a}{\sqrt{np_a(1-p_a)}} \right| < K,$$

otherwise **atypical**.
The set of $(n, \mathbf{p}, K)$-typical words denoted by

$$\mathbb{T}_{n,\mathbf{p},K} := \{\boldsymbol{\alpha} \in \mathbb{A}^n : \boldsymbol{\alpha} \quad \text{is } (n, \mathbf{p}, K)\text{-typical}\} \subset \mathbb{A}^n$$

Remark

If α typical word for vector $\mathbf{p}$, $\alpha \in \mathbb{A}^n$

$$\left| \frac{\nu_a^{(n)}(\boldsymbol{\alpha})}{n} - p_a \right| < \frac{K}{\sqrt{p_a(1-p_a)}} \frac{1}{\sqrt{n}} = \mathcal{O}(n^{-1/2}), \forall a \in \mathbb{A}.$$

Typical words depend on probability vector $\mathbf{p}$ but they are not random themselves.
$\mathbb{T}_{n,\mathbf{p},K}$ contains n-letter words with preset[a] density of letters.

[a] Determined by $\mathbf{p}$.

(3) Entropy is ...
...the logarithmic ratio of typical over total configurations (cont'd)

Theorem (Asymptotic equirepartition property)

Let $\varepsilon \in]0,1[$ *and* $K > \lceil \sqrt{\frac{\mathrm{card}\mathbb{A}}{\varepsilon}} \rceil$. *For* $n \geq K$,

1. $\mathbb{P}(\mathbf{X}\restriction_n \notin \mathbb{T}_{n,\mathbf{p},K}) < \varepsilon$;
2. $\exists c > 0$ *such that* $\forall \boldsymbol{\alpha} \in \mathbb{T}_{n,\mathbf{p},K}$, *we have*
$$2^{-nH(\mathbf{p})-c\sqrt{n}} \leq \mathbb{P}(\mathbf{X}\restriction_n = \boldsymbol{\alpha}) \leq 2^{-nH(\mathbf{p})+c\sqrt{n}};$$
3. $\mathrm{card}(\mathbb{T}_{n,\mathbf{p},K}) = 2^{n(H(\mathbf{p})+\delta_n)}$, *with* $\lim_{n\to\infty} \delta_n = 0$.

1-C: Significance and proof.

1 USB key : 1 Gb $\simeq \underline{8 \times 10^9 \text{ bits}}$

$2^{10^9} = 10^{3\times 10^8}$

$\mathbb{A} = \{0,1\}$ $n = 1000$

$|\mathbb{A}^n| = 2^{1000}$

$\underline{p} = (0.2,\ 0.8)$ $H(\underline{p}) = 0.72$

$\quad p_0 \quad p_1$

Fix $\varepsilon = 5\% = 0.05$

$K = \sqrt{\frac{\text{card}\,\mathbb{A}}{\varepsilon}} = \sqrt{\frac{2}{0.05}} = 6.32$ $\Big]$ 79.6

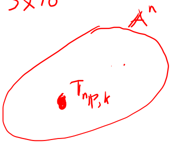
80

$\sqrt{n p_a (1-p_a)} = \sqrt{1000 \times 0.2 \times 0.8} = 12.6$

For $\alpha \in T_{n,p,K}$: $\nu_0(\underline{\alpha}) = 200 \pm \cancel{80}$

$\nu_1(\underline{\alpha}) = 800 \pm \cancel{80}$

a) $\frac{|T_{n,p,k}|}{|A^n|} \simeq \frac{2^{1000 \times 0.72}}{2^{1000}} = 2^{-280} = 5 \times 10^{-84}$



b) $X_1 \dots X_n$ iid $\sim P$

$P(\underline{X} \in T_{n,p,k}) \geq 95\%$

Although $T_{n,p,k}$ has ridiculously small cardinality it carries almost all proba. mass.

c) $\forall \underline{\alpha} \in T_{n,p,k} \quad P(\underline{X} = \underline{\alpha}) \simeq 2^{-720} = \frac{1}{\text{card } T_{n,p,k}}$.

AEP

Properties of entropy

Obviously $H(\mathbf{p}) \geq 0$ because $-p_a \log p_A \geq 0$ for all $a \in \mathbb{A}$.

Lemma

Let $\mathbf{p}, \mathbf{q} \in PV_{\mathrm{card}\mathbb{A}}$ *arbitrary probability vectors. Then*

$$-\sum_{a\in\mathbb{A}} p_a \log p_a \leq -\sum_{a\in\mathbb{A}} p_a \log q_a.$$

Proof.

- Function $t \mapsto \ln t$ is concave on $\mathbb{R}_+$.
- $\ln 1 = 0$ and $(\ln t)'|_{t=1} = \frac{1}{t}|_{t=1} = 1$.
- Concavity means that $\ln t \leq t - 1$ for all $t > 0$. Hence $\ln \frac{q_a}{p_a} \leq \frac{q_a}{p_a} - 1$ with equality iff $p_a = q_a$.
- We conclude

$$\sum_{a\in\mathbb{A}} p_a \ln \frac{q_a}{p_a} \leq \sum_{a\in\mathbb{A}} p_a\left(\frac{q_a}{p_a} - 1\right) = \sum_{a\in\mathbb{A}} (p_a - q_a) = 0.$$

□

Properties of entropy (cont'd)

Theorem

For every $\mathbf{p} \in PV_{\mathrm{card}\mathbb{A}}$,

$$H(\mathbf{p}) \leq \log \mathrm{card}\mathbb{A},$$

with equality iff $p_a = \frac{1}{\mathrm{card}\mathbb{A}}$ *for all* $a \in \mathbb{A}$.

Proof.

Apply previous lemma to $\mathbf{q}$ = uniform probability vector on $\mathbb{A}$, i.e. $q_a = \frac{1}{|\mathbb{A}|}$ for all $a \in \mathbb{A}$. Then $H(\mathbf{p}) \leq \log \mathrm{card}\mathbb{A}$ and bound is saturated iff $\mathbf{p} = \mathbf{q}$. □

Kullback-Leibler contrast
(or relative entropy)

- **p** and **q** probability vectors on same alphabet $\mathbb{A}$.
- **p** is **absolutely continuous** w.r.t. **q**, denote by $\mathbf{p} \ll \mathbf{q}$, if $p_a = 0$ for those $a \in \mathbb{A}$ for which $q_a = 0$, i.e. if $q_a = 0$ implies $p_a = 0$.

Definition

p, **q** probability vectors on same alphabet $\mathbb{A}$. **Relative entropy** or **Kullback-Leibler constrast** of **p** w.r.t. **q** the quantity

$$D(\mathbf{p}\|\mathbf{q}) := \begin{cases} \sum_{a\in\mathbb{A}} p_a \log\left(\frac{p_a}{q_a}\right) & \text{if } \mathbf{p} \ll \mathbf{q} \\ +\infty & \text{else.} \end{cases}$$

- $D(\mathbf{p}\|\mathbf{q}) \geq 0$, for arbitrary **p** and **q**.
- D is not symmetric in its arguments. Nevertheless, the larger the value of $D(\mathbf{p}\|\mathbf{q})$, the easier to discriminate between **p** and **q**.

Kullback-Leibler contrast (cont'd)

Coalescence increases entropy

Definition

Let $(\mathbb{X}, \mathbf{p})$ and $(\mathbb{Y}, \mathbf{q})$ be two probability spaces. $(\mathbb{Y}, \mathbf{q})$ is a **fragmentation of** $(\mathbb{X}, \mathbf{p})$ (or $(\mathbb{X}, \mathbf{p})$ is a **coalescence of** $(\mathbb{Y}, \mathbf{q})$) if can be partitioned into $\mathbb{Y} = \sqcup_{x\in\mathbb{X}} \mathbb{Y}_x$, so that for all $x \in \mathbb{X}$, $p(x) = \sum_{y\in\mathbb{Y}_x} q(y)$.

Proposition

For $i = 0, 1$, suppose $(\mathbb{Y}, \mathbf{q}_i)$ are fragmentations of $(\mathbb{X}, \mathbf{p}_i)$. Then

$$D(\mathbf{q_0}\|\mathbf{q_1}) \geq D(\mathbf{p_0}\|\mathbf{p_1}), \quad \textit{i.e. fragmentation increases Kullback-Leibler contrast.}$$

Proof.

W.l.o.g. suppose $\mathbf{q_0} \ll \mathbf{q_1}$.

$$\begin{aligned}
D(\mathbf{q_0}\|\mathbf{q_1}) - D(\mathbf{p_0}\|\mathbf{p_1}) &= \sum_{y\in\mathbb{Y}} q_0(y)\log\frac{q_0(y)}{q_1(y)} - \sum_{x\in\mathbb{X}} p_0(x)\log\frac{p_0(x)}{p_1(x)} \\
&= \sum_{x\in\mathbb{X}}\sum_{y\in\mathbb{Y}_x}\left(q_0(y)\log\frac{q_0(y)}{q_1(y)} - q_0(y)\log\frac{p_0(x)}{p_1(x)}\right) = \sum_{x\in\mathbb{X}}\sum_{y\in\mathbb{Y}_x} q_0(y)\log\frac{q_0(y)p_1(x)}{q_1(y)p_0(x)} \\
&\geq \sum_{x\in\mathbb{X}}\sum_{y\in\mathbb{Y}_x}\left(q_0(y) - q_0(y)\frac{q_1(y)p_0(x)}{q_0(y)p_1(x)}\right) \quad (\text{because } \log t \geq 1 - \frac{1}{t}) \\
&= 0.
\end{aligned}$$

$\square$

Kullback-Leibler contrast (cont(d)

Contrast of Markovian evolutions

Theorem

- $(X_n)_{n\in\mathbb{N}}$ *irreducible and aperiodic Markov chain on* ~~*denumerable*~~ finite *space* $\mathbb{X}$ *of stochastic matrix* P.
- π *its equilibrium probability:* $\pi = \pi P$.
- $\mu_n(y) := \mathbb{P}_\rho(X_t = y)$ *for arbitrary (fixed) initial probability* $\rho \in \cancel{\mathcal{M}_1(\mathbb{X})}$. $\mathrm{PV}_{\mathbb{X}}$
- $f : \mathbb{R}_+ \to \mathbb{R}$ *strictly concave measurable function and, for* $n \in \mathbb{N}$,

$$F_n = \sum_{y\in\mathbb{X}} \pi(y) f\left(\frac{\mu_n(y)}{\pi(y)}\right).$$

Under previous conditions, (F_n) *strictly increasing in* n.

Corollary

Under same conditions, $D(\mu_n\|\pi)$ *strictly decreasing in* n. *Tends to 0 when* $\mu_n \to \pi$.

1-D: Check whether basic notions on Markov chains are known.
1-E: Proof of theorem and corollary.

Markov chains on $\underline{\mathcal{X} \text{ finite}}$

$(X_n)_{n \in \mathbb{N}}$ sequence of $\mathcal{X}$-valued variables

"time"

$$\mathbb{P}(X_{n+1} = y \mid \underline{X_n = x_n, \dots X_1 = x_1})$$
$$= \mathbb{P}(X_{n+1} = y \mid X_n = x_n)$$

MARKOV PROP.

$$P(x,y) = \mathbb{P}(X_{n+1} = y \mid X_n = x)$$

$P = (P_{xy})_{x,y}$ $\quad |\mathcal{X}| \times |\mathcal{X}|$ matrix with elements in $[0,1]$

$$\sum_y P_{xy} = 1 \quad \forall x$$
$$\|$$
$$\sum_y \mathbb{P}(X_{n+1} = y \mid X_n = x) = 1$$

$P \longleftrightarrow$ directed graph

$L = $ fortune A + fortune B

Gambler's ruin.

Def: If $\forall (x,y) \in \mathcal{X}^2$, $\exists n = n(x,y)$: $(P^n)_{xy} > 0$

$\Rightarrow$ irreducible

strongly irred : if $\exists N$, $\exists \alpha > 0$: $\min_{x,y} (P^N)_{xy} = \alpha > 0$

Def: (X_n) MC.

A generalized r.v. $T \in \mathbb{N} \cup \{+\infty\}$ is a stopping time if $\forall n \in \mathbb{N}$, event $\{T = n\}$ completely determined by $X_0, \dots X_n$ (does not depend on $X_{n+1}, \dots$)

eg: $\cancel{X}_n$ temperature day n of year

$T = \inf\{n \geq 1 : X_n \geq 42.2\}$ $\qquad \inf \emptyset = +\infty$

$T' = \inf\{n \geq 1 : X_n = \text{max of the year}\}$

T is a stopping time but T' is not.

$(X_n)_{n \in \mathbb{N}}$ on $\mathcal{X}$ <u>finite</u> alphabet

$$\mathbb{P}(X_{n+1} = y \mid X_n = x_n, \dots X_0 = x_0) = \mathbb{P}(X_{n+1} = y \mid X_n = x_n)$$

$$\mathbb{P}(X_{n+1} = y \mid X_n = x) = P(x,y) \geq 0 \quad x, y \in \mathcal{X}$$

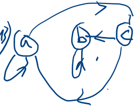
$$\forall x, \ \sum_y P(x,y) = 1$$

stochastic matrix

d) Relation R on X is specific $\subseteq X \times X$

$X = \{a,b,c\}$ $\quad \{(a,a)\ (a,c),\ (b,b)\ (c,a),\ (c,b)\} \subseteq X \times X$

$$M = \begin{array}{c|ccc} & a & b & c \\ \hline a & 1 & & 1 \\ b & & 1 & \\ c & 1 & 1 & \end{array} \in M_{3\times 3}(\{0,1\})$$



$dom(R) = \{ x \in X : \exists y \in X \ (x,y) \in R \}$

b) function special case relation

st. $\forall x$, at most one y : $(x,y) \in f$.

$x \in dom(f) \Rightarrow \forall x, \exists! y = y_x . (x,y_x) \in f$

c) stochastic Matrix ; $M \in M_{3\times 3}([0,1])$ $\quad \sum_y M(x,y) = 1 \ \forall x$

$(M(x,\cdot))$

$\mu_0 =$ proba $\sim X_0$ $\quad$ $\mu_0(x) = \mathbb{P}(X_0 = x)$

μ_1 $\quad$ $\mu_1(y) = \mathbb{P}_{\mu_0}(X_1 = y) = \sum_x \mu_0(x) P(x,y) = (\mu_0 P)(y)$

$\mu_n(y) = \mathbb{P}_{\mu_0}(X_n = y) = \sum_{x_0 \dots x_{n-1}} \mu_0(x_0) P(x_0 x_1) \dots P(x_{n-1}, y)$

$= (\mu_0 P^n)_y$

Def If $\mu_n \to \mu$ then μ is limiting distribution

Def. π proba vector : $\pi = \pi P$ left eigenvector of P ~ eigenvalue 1

invariant or stationary proba

Prop: (X_n) strongly irreducible and aperiodic and μ limiting then $\mu = \pi$

$d_x = \gcd\{n \geq 1 : (P^n)_{xx} > 0\}$ state x aperiodic $\Rightarrow d_x = 1$

Pf: $\mu_n = \mu_0 P^n \Rightarrow \mu_n = \mu_{n-1} P$ $\mu = \mu P$

$$\begin{array}{ccc} \mu_n & = & \mu_{n-1} P \\ {\scriptstyle n\to\infty}\downarrow & & \downarrow \\ \mu & & \mu \end{array}$$

$\mu = \pi$

Chain (process (X_n))

stationary : $\forall T$: joint proba of $(X_{T+1}, \dots X_{T+n}) =$ ~~law~~

$=$ joint proba of $(X_1, \dots X_n)$

reversible if law $(X_1 \dots X_n) =$ law $(X_{T-1} \dots X_{T-n})$

Lemme : Reversible $\Longrightarrow$ stationarity

$$\mathbb{P}(X_1 = x_1, \dots X_n = x_n) = K(x_1, \dots x_n)$$

$$K \in \mathrm{Pr}_{\mathcal{X}^n}$$

If (X_n) $MC(\mathcal{X}, P, \mu_0) =$ $\mu_0(x_1)\, P(x_1, x_2) \dots P(x_{n-1}, x_n)$

Other derived quantities
Joint entropy

- X and Y random variables on $\cancel{(\Omega, \mathcal{F}, \mathbb{P})}$ taking values in $\mathbb{X}$ and $\mathbb{Y}$,
- joint law of (X, Y) détermined by vector $\boldsymbol{\kappa}$ of joint probability: $\kappa(x, y) := \mathbb{P}(X = x; Y = y)$, for $x \in \mathbb{X}$ and $y \in \mathbb{Y}$.

Definition

X and Y as above with joint law $\boldsymbol{\kappa}$. **Joint entropy**

$$H(X, Y) := H(\boldsymbol{\kappa}) = -\sum_{(x,y)\in\mathbb{X}\times\mathbb{Y}} \underline{\kappa(x, y) \log \kappa(x, y)} = -\mathbb{E}(\log \kappa(X, Y)).$$

Similairly, if $\mathbf{X} = (X_1, \ldots, X_n)$ is a collection of random varaibles with joint law $\boldsymbol{\kappa}$, i.e. $\kappa(\mathbf{x}) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)$, the joint entropy reads

$$H(X_1, \ldots, X_n) = -\sum_{\mathbf{x}\in\mathbb{X}_1\times\cdots\times\mathbb{X}_n} \mathbb{P}(\mathbf{X} = \mathbf{x}) \log \mathbb{P}(\mathbf{X} = \mathbf{x}) = -\mathbb{E}(\log \kappa(\mathbf{X})).$$

Theorem

Joint entropy is sub-additive, i.e.

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

with equality iff X and Y independent.

1-F: Comment on theorem and proof of sub-additivity.

Subadditivity of $(a_n) \geq 0$

$$a_{n+m} \leq a_n + a_m \implies \lim \frac{a_n}{n} \text{ exists} = \inf \frac{a_n}{n}$$

$$H(X,Y) \overset{?}{\leq} H(X) + H(Y)$$

$$-\sum p_a \log p_a \leq -\sum p_a \log q_a$$

$$\kappa(x,y) = \mathbb{P}(X=x, Y=y)$$

$$\mu(x) = \sum_y \kappa(x,y) \qquad \nu(y) = \sum_x \kappa(x,y)$$

Let Z_1, Z_2 be indep r.v. $\quad Z_1 \sim \mu \quad Z_2 \sim \nu$

$$q(x,y) = \mu(x)\nu(y) =$$

$$H(X) = -\sum_x \mu(x) \log \mu(x)$$

$$H(Y) = -\sum_y \nu(y) \log \nu(y)$$

$$H(X,Y) = -\sum_{x,y} \kappa(x,y) \log \kappa(x,y) \leq -\sum_{x,y} \kappa(x,y) \log q(x,y)$$

Other derived quantities
Conditional entropy

- X, Y random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ with values in $\mathbb{X}$ in $\mathbb{Y}$.
- Knowing that $X = x$ occurred determine conditional law $\mathbb{P}(Y \in B | X = x)$, for $B \subseteq \mathbb{Y}$ and $x \in \mathbb{X}$.
- Conditional probability is a probability. Hence define the entropy of the conditional law by

$$H(Y|X = x) = -\sum_{y \in \mathbb{Y}} \mathbb{P}(Y = y|X = x) \log \mathbb{P}(Y = y|X = x).$$

Definition

Conditional entropy of Y given X = the average of entropies of the conditional laws, i.e.

$$H(Y|X) = \sum_{x \in \mathbb{X}} H(Y|X = x)\mathbb{P}(X = x).$$

Other derived quantities
Conditional entropy (cont'd)

Theorem

- $H(Y|X) = H(X,Y) - H(Y)$.
- $H(Y|X) \leq H(Y)$, *with equality iff* Y *and* X *independent.*

Proof.

$$\begin{aligned} H(Y|X) &= -\sum_{x\in\mathbb{X}} \mathbb{P}(X=x) \sum_{y\in\mathbb{Y}} \mathbb{P}(Y=y|X=x) \log \mathbb{P}(Y=y|X=x) \\ &= -\sum_{(x,y)\in\mathbb{X}\times\mathbb{Y}} \mathbb{P}(X=x, Y=y)[\log \mathbb{P}(Y=y; X=x) - \log \mathbb{P}(X=x)] \\ &= H(X,Y) - H(X). \end{aligned}$$

$$H(X,Y) = H(Y|X) + H(X) \overset{\text{subadditivity}}{\leq} H(Y) + H(X),$$

with equality iff Y and X independent. □

Other derived quantities

Mutual information

Definition

Mutual information of X and Y,

$$I(X:Y) := H(X) - H(X|Y).$$

Remark

- $H(X)$ a priori uncertainty on X (knowing only its law).
- $H(X|Y)$ residual uncetainty on X given that Y has been observed.
- $I(X:Y) := H(X) - H(X|Y)$ information on X mediated by observation of Y.
- From
$$H(X,Y) = H(Y|X) + H(X) = H(X|Y) + H(Y),$$
we conclude symmetry of mutual information:
$$\begin{aligned} I(X:Y) &= H(X) - H(X|Y) \\ &= H(X) + H(Y) - H(X,Y) = I(Y:X). \end{aligned}$$
- Mutual information useful to
  - define channel capacity (lecture 3),
  - prove unconditional security of certain cryptographic schemes.

Exercise

Theorem

Let $\mathbb{X}$ be finite, $\beta > 0$ real parameter, and $U : \mathbb{X} \to \mathbb{R}_+$. For arbitrary probability vector ν on $\mathbb{X}$, denote by $\nu U := \sum_{x \in \mathbb{X}} \nu(x) U(x)$, the expectation of U under ν. Then,

1. *there exists probability vector μ_β on $\mathbb{X}$ saturating the $\sup_\nu (H(\nu) - \beta \nu U)$, where $H(\nu)$ is the entropy of ν,*
2. *$\mu_\beta(x) = \frac{\exp(-\beta U(x))}{\mathcal{Z}(\beta)}$, for $x \in \mathbb{X}$, where $\mathcal{Z}(\beta) = \sum_{y \in \mathbb{X}} \exp(-\beta U(y))$ is a normalising factor.*

Exercise

The purpose of this exercise is to prove the previous theorem.

1. Use concavity of log to show that for all probability vectors ν on $\mathbb{X}$, we have $H(\nu) - \beta \nu U \le \log \mathcal{Z}(\beta)$.
2. Compute $H(\mu_\beta) - \beta \mu_\beta U$.

a) $H(\nu) - \beta\nu U = - \sum \nu(x)\left(\log \nu(x) + \beta U(x)\right)$

$$= \sum \nu(x) \log \frac{e^{-\beta U(x)}}{\nu(x)}$$

$$\leq \log \sum \cancel{\nu(x)} \frac{e^{-\beta U(x)}}{\cancel{\nu(x)}}$$

$$= \log Z(\beta)$$

b) Compute $H(\mu_\beta) - \beta\mu_\beta U = \sum \mu_\beta(x)\left[\beta \cancel{U(x)} + \log Z(\beta) - \cancel{\beta U(x)}\right]$

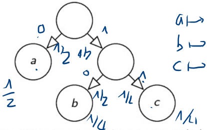
$$= \log Z(\beta)$$

Problem

Using a honest coin to simulate a biased one

- We have honest coin (taking values in $\mathbb{B} = \{0, 1\}$ with probability vector $\mathbf{p} = (1/2, 1/2)$.) Successive tosses: sequences of arbitrary lengthof random bits $\xi \in \mathbb{B}^+$.
- Want to simulate random variable X on finite set $\mathbb{X}$, distributed according to $\mathbf{p} := (p(x))_{x \in \mathbb{X}}$.

1. Start with set $\mathbb{X} = \{a, b, c\}$ and $\mathbf{p} = (1/2, 1/4, 1/4)$. Place letters a, b, c on leaves of complete binary tree (i.e. every node has either 0 or 2 children) as on adjacent figure. Associate bit 0 with left edges and 1 with right ones. Observe that set of leaves $\mathbb{F} = \{0, 10, 11\}$ can be surjected into $\mathbb{X}$. Denote $F : \mathbb{F} \to \mathbb{X}$ this surjection (in the present special case, F is a bijection.) Give explicit algorithm of generating X.



$a \mapsto 0$
$b \mapsto 10$
$c \mapsto 11$

2. In this special case, estimate mean number of tosses to simulate X; compare with entropy $H(X)$.
3. Let now $\mathbb{X} = \{a, b\}$ and $\mathbf{p} = (2/3, 1/3)$. Use representation of $2/3$ and $1/3$ to determine set of leaves $\mathbb{F}$ and surjection $F : \mathbb{F} \to \mathbb{X}$. *Hint:* $\sum_{k=0}^{\infty} \frac{1}{2^{2k+1}} = \frac{2}{3}$.

$\frac{2}{3} = \langle 0.1010101\ldots\rangle_2$

4. Estimate mean number of tosses to simulate X and compare with entropy. *Hint:* $\sum_{k=0}^{\infty} \frac{k}{2^k} = 2$.