III. Data compression

III.1. Symbol codes

Motivation: in natural languages letters come with unequal probabilities (e.g. Hawaiian $p(\text{"a"}) \gtrsim 26\%$ whereas in Swedish any letter $\leq 10.5\%$)

→ Texts can be compressed by encoding them so that more likely letters are represented by shorter binary strings ...

Def.:

- $A^+ := \bigcup_{n \in \mathbb{N}} A^{\times n}$ = set of finite strings/words $a_1 \dots a_n$ with $a_i \in A, n \in \mathbb{N}$
- $A^* = A^+ \cup \{\varepsilon\}$ (ε = empty word)
- Let X be a random variable with range $\mathcal{X}$. A "symbol code" for X is a map $C: \mathcal{X} \to A^+$. $C(\mathcal{X})$ is the set of "codewords", $l_x \in \mathbb{N}$ the respective length and $L(C) := \sum_{x \in \mathcal{X}} p(x) l_x$ the average length.
- A symbol code C is called

  (i) "non-singular" iff C is injective,

  ⇑

  (ii) "uniquely decodable" iff the map $\mathcal{X}^+ \to A^+$ defined by $x_1 \dots x_n \mapsto C(x_1) \dots C(x_n)$ is injective,

  ⇑

  (iii) "prefix-free" (or sometimes a "prefix code") iff there is no pair $x \neq x'$ s.t. $C(x) = C(x')a$ for some $a \in A^*$

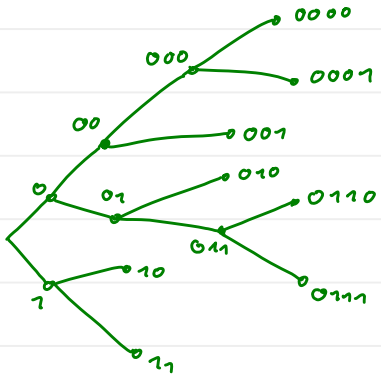
Examples:

- non-singular but not uniquely decodable:
  $x_1 \mapsto 0,\ x_2 \mapsto 1,\ x_3 \mapsto 01$
- uniquely decodable but not prefix-free:
  $x_1 \mapsto 0,\ x_2 \mapsto 01$
- prefix-free: $x_1 \mapsto 0,\ x_2 \mapsto 10,\ x_3 \mapsto 110$

- ASCII is prefix-free since all codewords have equal length
- Morse code is not prefix-free (overcome by time gaps between letters)

Code trees:

A "code tree" for a prefix-free code is an $|A|$-ary tree of depth $\max\{l_x\}_{x\in\mathcal{X}}$, so that $\{\text{leaves}\} = \{\text{codewords}\}$



(for non-prefix free codes intermediate vertices could also correspond to codewords)

Thm.: (Kraft's inequality)

- For any prefix-free code $C: \mathcal{X} \to A^+$ with codeword lengths $l_1, \dots, l_{|\mathcal{X}|}$

$$\boxed{\sum_{x\in\mathcal{X}} |A|^{-l_x} \leq 1} \quad \text{holds.}$$

- Conversely, given $l \in \mathbb{N}^{|\mathcal{X}|}$ satisfying this inequality there exists a prefix-free code with those codeword lengths.

proof: $l_m := \max\{l_x\}_{x\in\mathcal{X}}$. Consider the $|A|$-ary tree of depth l_m with $|A|^{l_m}$ leaves.

If a codeword has length l_x, this reduces the number of leaves by at least $\left(|A|^{l_m - l_x} - 1\right)$.

$$\rightarrow \quad |\mathcal{X}| \leq |A|^{l_m} - \left(\sum_{x\in\mathcal{X}} |A|^{l_m - l_x} - 1\right)$$

($|\mathcal{X}|$: total # of leaves)

$$\Leftrightarrow \quad \sum_{x\in\mathcal{X}} |A|^{-l_x} \leq 1$$

Converse: shown by explicit construction via code tree ... □

Generalization due to McMillan:

Thm.: Let $\mathcal{X}$ be countable infinite. Every uniquely decodable code satisfies Kraft's inequality.
Conversely, if $-"-$ holds for some $L \in \mathbb{N}^{|\mathcal{X}|}$, then there exists a prefix-free code with these codeword lengths $(L_x)_{x \in \mathcal{X}}$.

Def.: A symbol code $C: \mathcal{X} \to A^+$ is called "optimal" w.r.t. a random variable X iff the average codeword length $L(C)$ is minimal among all symbol codes in $(A^+)^{\mathcal{X}}$.

Thm.: (entropy bound for optimal codes)
Let $C: \mathcal{X} \to A^+$ be a uniquely decodable code for a random variable X.

Then $$\boxed{L(C) \geq \frac{H(X)}{\log |A|}}$$ with equality iff $L_x = -\log_{|A|} p(x) \quad \forall x \in \mathcal{X}$.

Conversely, there exists a prefix-free code for which:

$$\boxed{L(C) < \frac{H(X)}{\log |A|} + 1}$$

proof: ○ lower bound:

$$L(C) - \frac{H(X)}{\log |A|} = \sum_{x \in \mathcal{X}} p(x) \left(L_x - \log_{|A|} \frac{1}{p(x)} \right)$$

$$= \sum_{x \in \mathcal{X}} p(x) \left(\log_{|A|} p(x) - \log_{|A|} \left(|A|^{-L_x} \right) \right)$$

$$q(x) := \frac{|A|^{-L_x}}{\sum_{y \in \mathcal{X}} |A|^{-L_y}}$$

$$= \left(\sum_{x \in \mathcal{X}} p(x) \log_{|A|} \frac{p(x)}{q(x)} \right) - \log_{|A|} \sum_{y \in \mathcal{X}} |A|^{-L_y}$$

$= D(p \| q) \geq 0$ (Gibbs' inequality)

Kraft: ≤ 1

- equality requires

(i) $D(p\|q) = 0 \iff \forall x: p(x) = q(x)$

and (ii) $\sum_{x \in \mathcal{X}} |A|^{-l_x} = 1$

which is equivalent to $p(x) = q(x) = |A|^{-l_x} \quad \forall x \in \mathcal{X}$

and thus $l_x = \log_{|A|} \frac{1}{p(x)} \quad \forall x \in \mathcal{X}$

- upper bound: define a set of integers $l_x := \lceil \log_{|A|} \frac{1}{p(x)} \rceil$.

Then $l \in \mathbb{N}^{|\mathcal{X}|}$ satisfies Kraft's inequality since

$$\sum_{x \in \mathcal{X}} |A|^{-l_x} \leq \sum_{x \in \mathcal{X}} |A|^{\log_{|A|} p(x)} = 1$$

$\rightarrow$ there exists a prefix-free code with codeword lengths l.

$$L(C) = \sum_{x \in \mathcal{X}} p(x) \lceil \log_{|A|} \frac{1}{p(x)} \rceil < \sum_{x \in \mathcal{X}} p(x) \left(1 + \log_{|A|} \frac{1}{p(x)}\right)$$

$$= 1 + \frac{H(X)}{\log |A|} \qquad \square$$

remark: the resulting code is called "Shannon code"

Thm.: (entropy rate bounds)

Let $\{X_i\}_{i \in \mathbb{N}}$ be a stationary stochastic process. Then

(i) for all $N \in \mathbb{N}$ and all uniquely decodable codes $C: \mathcal{X}^N \to \{0,1\}^+$:

$$\boxed{H(\{X_i\}) \leq \frac{1}{N} L(C)}$$

(ii) $\forall \varepsilon > 0 \; \exists N \in \mathbb{N}$ and a prefix-free code $C: \mathcal{X}^N \to \{0,1\}^+$ such that

$$\boxed{\frac{1}{N} L(C) < H(\{X_i\}) + \varepsilon}$$

proof: (i) $H(\{X_i\}) \leq \frac{1}{N} H(X_1, \ldots, X_N) \leq \frac{1}{N} L(C)$

↑ non-increasing sequence due to stationarity

↑ entropy bound

(ii) $\forall \delta > 0 \; \exists N \in \mathbb{N} : \frac{1}{N} H(X_1, \ldots, X_N) < H(\{X_i\}) + \delta$

↑ convergence towards entropy rate due to stationarity

$\exists C : \mathcal{X}^N \to \{0,1\}^+$ such that

$$\frac{1}{N} L(C) < \frac{1}{N} H(X_1, \ldots, X_N) + \frac{1}{N} < H(\{X_i\}) + \underbrace{\delta + \frac{1}{N}}_{\varepsilon}$$

↑ entropy bound

□