X_i i.i.d. $\sim p(x)$ (PMF)

$$-\frac{1}{n}\log p^{(n)}(X_1,\dots,X_n) = -\frac{1}{n}\sum_{i=1}^{n}\log p(X_i)$$

$$\longrightarrow \mathbb{E}[-\log p(X_1)] = H(X)$$

(a.e., ~~abs~~ convergence in prob.)

$$A_\varepsilon^{(n)} = \left\{(x_1,\dots,x_n)\in \mathcal{X}^n \;\middle|\; 2^{-n(H(X)+\varepsilon)} \le p^{(n)}(x_1,\dots,x_n) \le 2^{-n(H(X)-\varepsilon)}\right\}$$

typical set.

Th. 2.4.4.

a) If $(x_1,\dots,x_n)\in A_\varepsilon^{(n)}$ then

$$H(X)-\varepsilon \le -\frac{1}{n}\log p^{(n)}(x_1,\dots,x_n) \le H(X)+\varepsilon$$

b) $P((X_1,\dots,X_n)\in A_\varepsilon^{(n)}) > 1-\varepsilon$ for n sufficiently large

c) $|A_\varepsilon^{(n)}| \le 2^{n(H(X)+\varepsilon)}$ ($|\cdot|$ cardinality)

d) $|A_\varepsilon^{(n)}| \ge (1-\varepsilon)\,2^{n(H(X)-\varepsilon)}$ for n suff. large.

Proof. a) ✓ b) ✓

c) $$1 = \sum_{x\in\mathcal{X}^n} p^{(n)}(x) \ge \sum_{x\in A_\varepsilon^{(n)}} p^{(n)}(x) \ge \sum_{x\in A_\varepsilon^{(n)}} 2^{-n(H(X)+\varepsilon)}$$

$$= 2^{-n(H(X)+\varepsilon)}\,|A_\varepsilon^{(n)}|$$

d) For n sufficiently large

$$1-\varepsilon < P((X_1,\dots,X_n)\in A_\varepsilon^{(n)}) \le \sum_{x\in A_\varepsilon^{(n)}} 2^{-n(H(X)-\varepsilon)}$$

$$= |A_\varepsilon^{(n)}|\,2^{-n(H(X)-\varepsilon)}$$

∎

For given $\varepsilon > 0$ and sufficiently large n

$\mathcal{X}^n$ decomposes into a set $T = A_\varepsilon^{(n)}$ (typical set)

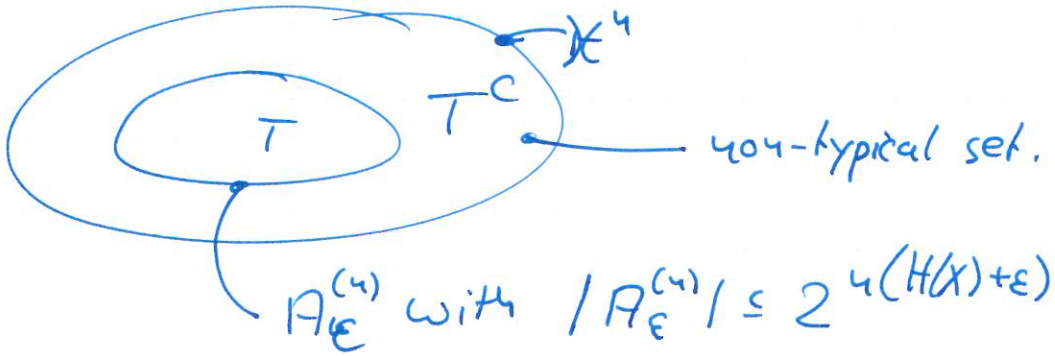
such that

- $P((X_1, \dots, X_n) \in T^c) \leq \varepsilon$
- For all $x = (x_1, \dots, x_n) \in T$:

$$\left| -\frac{1}{n} \log p^{(n)}(x_1, \dots, x_n) - H(X) \right| \leq \varepsilon$$

the normalized log-prob of all sequences in T is nearly equal and close to $H(X)$.

Graphically:



The AEP and Data Compression

Let $X_1, \dots, X_n$ i.i.d. with support $\mathcal{X}$, $X^{(n)} = (X_1, \dots, X_n)$.

Aim: find a short description/encoding of all values $x^{(n)} = (x_1, \dots, x_n) \in \mathcal{X}^n$.

Key idea: index coding, allocate each of the $|\mathcal{X}^n|$ values an index.

If holds • $|A_\varepsilon^{(n)}| \leq 2^{n(H(X)+\varepsilon)}$ (Th. 2.4.4 c))

Indexing of all $x^{(n)} \in A_\varepsilon^{(n)}$ requires at most $n(H(X)+\varepsilon)+1$ (1 bit extra since $n(H(X)+\varepsilon)$ may not be integer.)

• $|\mathcal{X}^n|$ requires $n \log|\mathcal{X}| + 1$ bits as indices.

Prefix each codeword for $x^{(n)} \in A_\varepsilon^{(n)}$ by 0 and each codeword for $x^{(n)} \notin A_\varepsilon^{(n)}$ by 1.

Let $\ell(x^{(n)})$ denotes the length of the codeword for $x^{(n)}$.

Then

$$\mathbb{E}[\ell(X^{(n)})] = \sum_{x^{(n)} \in \mathcal{X}^n} p(x^{(n)})\, \ell(x^{(n)})$$

$$= \sum_{x^{(n)} \in A_\varepsilon^{(n)}} p(x^{(n)})\, \ell(x^{(n)}) + \sum_{x^{(n)} \notin A_\varepsilon^{(n)}} p(x^{(n)})\, \ell(x^{(n)})$$

$$\leq \sum_{x^{(n)} \in A_\varepsilon^{(n)}} p(x^{(n)}) \left(n(H(X)+\varepsilon)+2\right)$$

$$+ \sum_{x^{(n)} \notin A_\varepsilon^{(n)}} p(x^{(n)}) \left(n \log|\mathcal{X}| + 2\right)$$

$$= P(X^{(n)} \in A_\varepsilon^{(n)}) \left(n(H(X)+\varepsilon)+2\right) + P(X^{(n)} \notin A_\varepsilon^{(n)}) \left(n \log|\mathcal{X}|+2\right)$$

$$\leq n(H(X)+\varepsilon) + \varepsilon n \log|\mathcal{X}| + 2$$

$$= n\Big(H(X) + \underbrace{\varepsilon \cancel{+\frac{1}{n}} + \varepsilon \log|\mathcal{X}| + \frac{2}{n}}_{\varepsilon'}\Big)$$

$$= n(H(X)+\varepsilon') \quad \text{for any } \varepsilon' > 0 \text{ with } n \text{ suff. large.}$$

It follows:

Th. 2.4.5. $\{X_n\}$ i.i.d. For any $\varepsilon > 0$ there exists $n \in \mathbb{N}$ and a binary code that maps each $x^{(n)}$ one-to-one onto a binary string satisfying

$$E\left(\frac{1}{n}\ell(X^{(n)})\right) \leq H(X)+\varepsilon. \quad \square$$

Hence, for suff. large n there exists a code for $X^{(n)}$ such that the expected average codeword length is arbitrary close to $H(X)$.

2.5. Differential Entropy

By now: Entropy for discrete r.v. with finite support.

Extension: discrete r.v. but countably many support points, $\mathcal{X} = \{x_1, x_2, \dots\}$ and distr. $\in p = (p_1, p_2, \dots)$

$$H(X) = -\sum_{i=1}^{\infty} p_i \log p_i .$$

Note: The sum may be infinite or may not even exist.

Important: Extension of entropy to r.v. X with a density f.

Def. 2.5.1. Let X be abs.-continuous with density $f(x)$.

$$h(X) = -\int_{-\infty}^{\infty} f(x) \log f(x)\, dx$$

is called differential entropy of X. ⊥

Remarks.

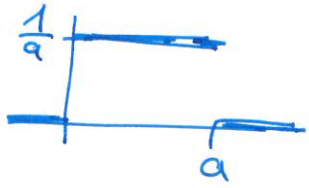
a) The integral in Def 2.5.1 may be infinite or may not even exist. (→ Exercises)

b) As a general implicit assumption in defining $h(x)$ we include: "provided the integral exists".

c) $h(X) = E[-\log f(X)]$

Examples 2.5.2

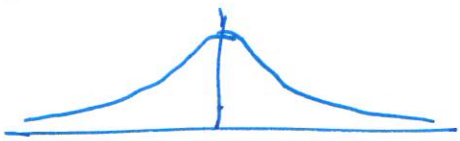
a) $X \sim U(0,a)$, $f(x) = \frac{1}{a} \mathbb{1}_{(0,a]}(x)$



$$h(X) = -\int_0^a \frac{1}{a} \log \frac{1}{a} \, dx$$

$$= \log a \quad , a > 0$$

b) $X \sim N(\mu, \sigma^2)$, $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, $x \in \mathbb{R}$



$$h(X) = \frac{1}{2} \ln(2\pi e \sigma^2)$$ Ex

Def. 2.5.3.

a) $X = (X_1, \dots, X_n)$ a r.vector with joint density $f(x_1, \dots, x_n)$.

$$h(X_1, \dots, X_n) = -\int \cdots \int f(x_1, \dots, x_n) \log f(x_1, \dots, x_n) \, dx_1 \cdots dx_n$$

is called joint ~~entropy~~ differential entropy of X.

b) (X, Y) a r.vector with joint density $f(x,y)$ and conditional density

$$f(x|y) = \frac{f(x,y)}{f(y)}, \text{ if } f(y) > 0,$$

and 0 otherwise.

$$h(X|Y) = -\int\int f(x,y) \log f(x|y) \, dx \, dy$$

is called conditional diff. entropy of X given Y.

Example 2.5.4 (n-dim Gaußian distr., $X \sim N_n(\mu, C)$)

$\mu \in \mathbb{R}^n$, $C \in \mathbb{R}^{n\times n}_{\text{sym, pos.def.}}$

$$f(x) = \frac{1}{(2\pi)^{n/2}\,|C|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T C^{-1}(x-\mu)\right\},$$

$x = (x_1, \dots, x_n)^T \in \mathbb{R}^n$.

$(|\cdot| = |\det(\cdot)|)$

$$h(X) = \frac{1}{2}\ln\left((2\pi e)^n |C|\right) \qquad \text{(Ex.)}$$

Def. 2.5.5. The mutual information between two r.v. X and Y with joint density $f(x,y)$ is defined as

$$I(X;Y) = h(X) - h(X|Y) = h(Y) - h(Y|X)$$

Interpretation: Amount of information about X from Y and vice versa.