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

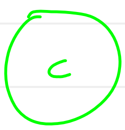
  $p : \mathcal{X} \to \mathbb{R}_+ := [0,\infty)$

  $\sum_{x\in\mathcal{X}} p(x) = 1$ (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 $p_X(x) := \sum_{y\in\mathcal{Y}} p(x,y)$, $p_Y(y) := \sum_{x\in\mathcal{X}} p(x,y)$
- conditional distribution $p(x|y) := \dfrac{p(x,y)}{p_Y(y)}$ for $p_Y(y) \neq 0$
- X & Y are independent r.v.s iff $p(x,y) = p_X(x)\, p_Y(y)$ $\forall x,y$

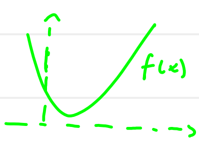
  $\Leftrightarrow p(x|y) = p_X(x)$ $\forall x\, \forall y : p_Y(y) > 0$

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) $\forall x \in C: f''(x) > 0 \implies f$ 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 ... $\square$

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.

- $\lceil x \rceil := \min\{n \in \mathbb{Z} \mid n \geq x\}$
- $\log x := \log_2 x$

- necessary: $2^n \geq |\mathcal{X}_0|$, $\mathcal{X}_0 := \{x \in \mathcal{X} \mid p(x) > 0\}$
- $\lceil \log |\mathcal{X}_0| \rceil$ 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)

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

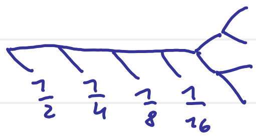
$$\xrightarrow{m \to \infty} \log|\mathcal{X}_0| =: H_0(X)$$ "Hartley entropy" or "0-Renyi entropy"

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



$\rightarrow$ 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 $\quad H(X) := -\sum_{x: p(x)>0} p(x) \log p(x)$

$$= -\sum_x p(x) \log p(x) \quad \text{with} \quad 0 \log 0 := 0$$

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

properties:
(i) symmetry: $\quad H(p_1, \dots, p_n) = H(p_{\pi(1)}, \dots, p_{\pi(n)}) \quad \forall \pi \in S_n$
(ii) expansibility: $\quad H(p_1, \dots, p_n, 0) = H(p_1, \dots, p_n)$
(iii) additivity: $\quad H(XY) = H(X) + H(Y)$ if X, Y independent
(iv) subadditivity: $\quad H(XY) \leq H(X) + H(Y)$

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

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

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

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

- Jensen
- $-\log$ convex

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

□

Thm.: (axiomatic characterization of entropies)

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

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

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

proof: [Aczél, Forte, Ng 1974]

Thm.: (bounds on Shannon entropy)

Let $p \in \mathbb{R}^n$ be a probability distribution and define $H_2(p) := -\log \|p\|_2^2$. Then

$$0 \overset{(i)}{\leq} H_2(p) \overset{(ii)}{\leq} H(p) \overset{(iii)}{\leq} H_0(p) \overset{(iv)}{\leq} \log n$$

where equality holds in

(i) iff $\exists x: p(x) = 1$

(ii) iff $\exists m \in \{1, \dots, n\}\ \forall x: p(x) \in \{0, \frac{1}{m}\}$

(iii) iff — " —

(iv) iff $\forall x: p(x) > 0$

proof: (i) $-\log \sum_x p(x)^2 \geq -\log \sum_x p_x = 0$

'=' iff $\forall x: p(x)^2 = p(x) \Leftrightarrow \forall x: p(x) \in \{0,1\}$

(ii) $H(p) = -\sum_x p(x) \log p(x) \geq -\log \sum_x p(x)^2$

$-\log$ is strictly convex

(iii) $H(p) = \sum_{x \in \mathcal{X}_0} p(x) \log \frac{1}{p(x)} \leq \log \sum_{x \in \mathcal{X}_0} \frac{p(x)}{p(x)} = H_0(p)$

log is strictly concave

(iv) ✓ □

II.2. Conditional entropy & mutual information

Def.:

- „conditional entropy" $H(X|Y) := H(X,Y) - H(Y)$
- „mutual information" $I(X:Y) := H(X) + H(Y) - H(X,Y)$
- „cond. mutual info." $I(X:Y|Z) := H(X,Z) + H(Z,Y) - H(X,Y,Z) - H(Z)$

Interpretation: • $H(X|Y) = \sum_y p(y) \underbrace{\left(\sum_x p(x|y) \log p(x|y)^{-1}\right)}_{\text{entropy of } X \text{ if } Y=y \text{ is known}}$

= average uncertainty about X if Y is known

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

= reduction of uncertainty (increase of information) about X after learning Y (and vice versa)

= information of X about Y and v.v.

• $I(X:Y|Z) = H(X|Z) - H(X|Z,Y)$

= reduction of uncertainty about X when we learn Y and already know Z.

Thm.: a) $H(X|Y) \geq 0$ with '=' iff $\forall y \exists x: p(x,y) = p(y)$

b) $I(X:Y) \geq 0$ with '=' iff $\forall x,y: p(x,y) = p(x)p(y)$

c) $I(X:Y|Z) \geq 0$

proof: a) $H(X,Y) - H(X) = \sum_{x,y} p(x,y) \log \underbrace{\left(\frac{\sum_{y'} p(x,y')}{p(x,y)}\right)}_{\geq 1} \geq 0$

"=" $\Leftrightarrow \forall x,y: p(x,y) = 0 \vee p(x) = p(x,y)$

b) $I(X:Y) = -\sum_{x,y} p(x,y) \log\left(\frac{p(x)p(y)}{p(x,y)}\right) \geq -\log \sum_{x,y} p(x)p(y) = 0$

$-\log$ strictly convex

'=' iff $\frac{p(x)p(y)}{p(x,y)} = \text{const.} = 1$

normalization

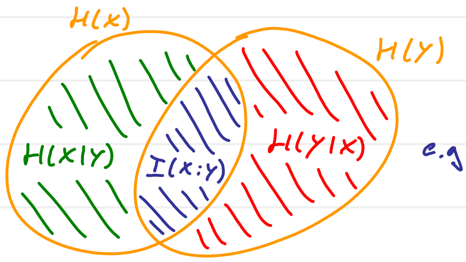
c) → exercise ... $\square$

remark: a) $\Leftrightarrow H(X,Y) \geq H(Y)$ can be seen as:

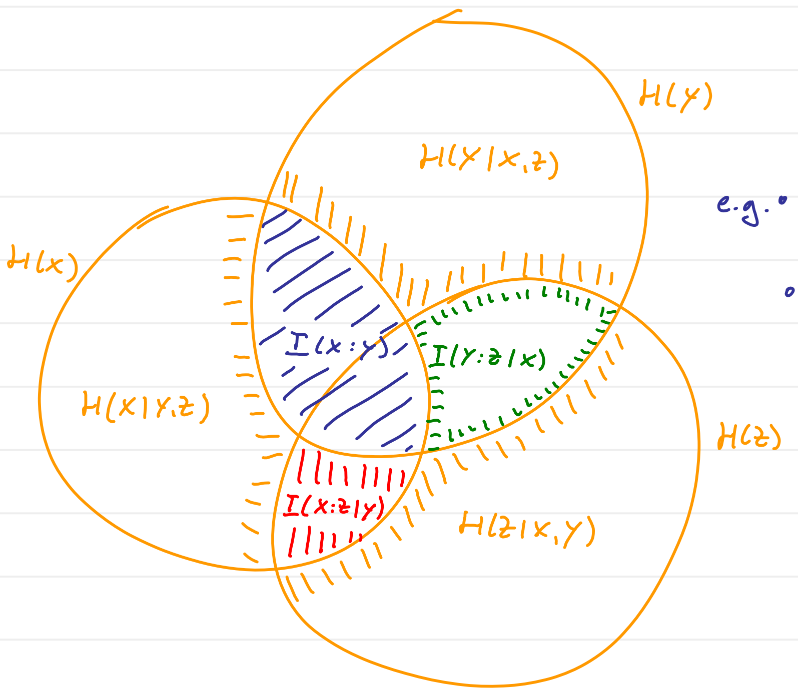
the entropy of a subsystem never exceeds the entropy of the whole system.

Venn-diagrams:

graphical depiction of relations between entropic quantities in terms of relations between sets:



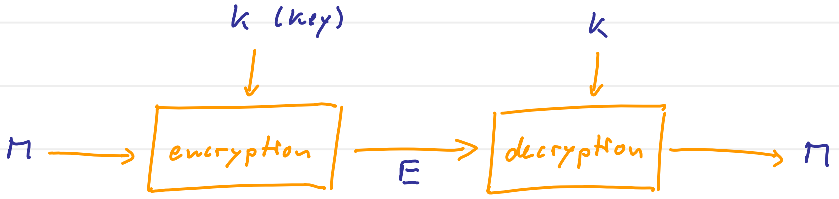
e.g. $I(X:Y) + H(Y|X) = H(Y)$



e.g. • $H(Y|X,Z) + I(Y:Z|X) = H(Y|X)$

• $H(X) = H(X|Y,Z) + I(X:Z|Y) + I(X:Y)$

II.3. Application for crypto systems



Def.: We say that random variables M, E, K describe a „perfectly secure crypto system" iff

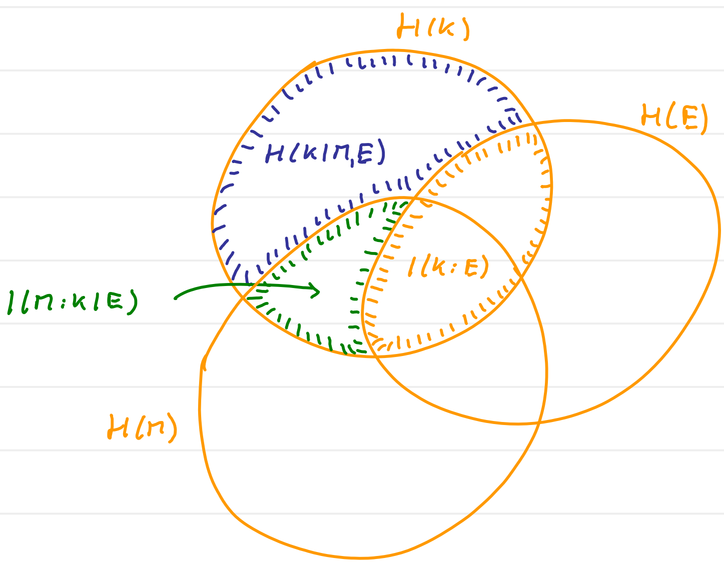
(i) $I(M:E) = 0$ (E contains no info about M without K)

(ii) $H(M|KE) = 0$ (once E and K are known, M can be perfectly recovered)

Thm.: (Shannon '49)

A perfectly secure crypto system requires $H(K) \geq H(M)$.

proof:
$$H(K) = I(M:K|E) + \underbrace{I(K:E)}_{\geq 0} + \underbrace{H(K|M,E)}_{\geq 0}$$
$$\geq I(M:K|E) = H(M) - \underbrace{H(M|K,E)}_{=0} - \underbrace{I(M:E)}_{=0} = H(M)$$
$\square$



remark:
- loosely speaking, this means that the key must not be shorter than the message.
- in practice, of course, weaker requirements are imposed.

II.4. Chain rules

Thm.: (a) $H(X_1, \dots, X_n) = \sum_{i=1}^{n} \underbrace{H(X_i | X_{i-1}, \dots, X_1)}_{:= H(X_1) \text{ for } i=1}$

(b) $H(X_1, \dots, X_n | Y) = \sum_{i=1}^{n} \underbrace{H(X_i | X_{i-1}, \dots, X_1, Y)}_{:= H(X_1 | Y) \text{ for } i=1}$

(c) $I(X_1, \dots, X_n ; Y) = \sum_{i=1}^{n} \underbrace{I(X_i ; Y | X_{i-1}, \dots, X_1)}_{:= I(X_1 ; Y) \text{ for } i=1}$

proof (sketch): for (a) use $p(x_1, \dots, x_n) = \prod_{i=1}^{n} \frac{p(x_i, \dots, x_1)}{p(x_{i-1}, \dots, x_1)} \leftarrow 1 \text{ for } i=1$

$$= \prod_{i=1}^{n} p(x_i | x_{i-1}, \dots, x_1)$$

und similarly for (b).

(c): $I(X_1, \dots, X_n ; Y) = H(X_1, \dots, X_n) - H(X_1, \dots, X_n | Y)$

$$\overset{(a),(b)}{=} \sum_{i=1}^{n} H(X_i | X_{i-1}, \dots, X_1) - H(X_i | X_{i-1}, \dots, X_1, Y)$$

$$= \sum_{i=1}^{n} I(X_i ; Y | X_{i-1}, \dots, X_1)$$

$\square$

II.5. Data processing inequality

Def.: • A "Markov chain" is a (finite or infinite) sequence of random variables $\{X_i\}_{i\in\mathbb{N}}$ for which

$$p(x_n | x_{n-1}, \dots, x_1) = p(x_n | x_{n-1}) \quad \text{for all } n \in \mathbb{N} \text{ and all } x\text{'s}.$$

• A Markov chain is called "stationary" or "homogeneous" iff $p(X_n = a | X_{n-1} = b) = p(X_2 = a | X_1 = b)$ for all n, a, b.

remarks: • Markov chains are often indicated by $X_1 \to X_2 \to X_3 \to \dots$ or, equivalently, $X_1 \leftarrow X_2 \leftarrow X_3 \leftarrow \dots$ (we will write $X_1 - X_2 - X_3 - \dots$)

• The probability distribution characterizing a Markov chain is

$$p(x_n, \dots, x_1) = p(x_n | x_{n-1}) \dots p(x_2 | x_1)\, p(x_1) \qquad X_1 \to X_2 \to \dots$$

$$= p(x_1 | x_2) \dots p(x_{n-1} | x_n)\, p(x_n) \qquad X_1 \leftarrow X_2 \leftarrow \dots$$

for $X \to Y \to Z$ this means $\quad p(x,y,z) = \dfrac{p(x,y)\,p(y,z)}{p(y)} \quad \forall x,y,z$

Prop.: X, Y, Z form a Markov chain $X - Y - Z$ iff $I(X:Z|Y) = 0$.

proof: $I(X:Z|Y) = \sum_{x,y,z} p(x,y,z) \log\left[\dfrac{p(x,y,z)\,p(y)}{p(x,y)\,p(y,z)}\right]$ $\square$

Lemma: If $Z = f(Y)$ for some $f: \mathbb{R} \to \mathbb{R}$, then $X - Y - Z$ is a Markov chain.

proof: $p(x|y,z) = p(x|y)$ $\square$

Thm.: ("data processing inequality")

If $X - Y - Z$ is a Markov chain, then $H(X|Y) \leq H(X|Z)$ and

$$\boxed{I(X:Y,Z) = I(X:Y) \geq I(X:Z)}$$

proof: using chain rules in two different ways we get:

$$I(X:Y,Z) = \begin{cases} I(X:Z) + \overbrace{I(X:Y|Z)}^{\geq 0} \\ I(X:Y) + \underbrace{I(X:Z|Y)}_{=0} \end{cases}$$

$\square$

interpretation: ∘ Z contains no more information about X than Y does.

∘ processing information (about X) cannot increase it.

II.6. Fano's inequality

Quantitative version of: "if Y allows to estimate X well, then $H(X|Y)$ is small."

For random variables X, Y we define $p_e := p(Y \neq X)$

$h(p_e) := H((p_e, 1-p_e))$ "binary entropy"

Thm.: (Fano's inequality) If X, Y are random variables with $\mathrm{range}(X) = \mathcal{X}$.
Then

$$\boxed{h(p_e) + p_e \log|\mathcal{X}| \geq H(X|Y)}$$

proof: define a random variable $E := \begin{cases} 1, \text{if } Y \neq X \\ 0 \end{cases}$

from the chain rule we obtain:

$$H(E,X|Y) = H(X|Y) + \overbrace{H(E|X,Y)}^{\overset{(i)}{=0}}$$
$$= \underbrace{H(E|Y)}_{\overset{(ii)}{\leq} h(p_e)} + \underbrace{H(X|E,Y)}_{\overset{(iii)}{\leq} p_e \log|\mathcal{X}|}$$

(i) E is a function of X and Y $\Rightarrow$ $H(E|X,Y) = 0$

(ii) $H(E|Y) \underset{\overset{\uparrow}{I(E;Y)\geq 0}}{\leq} H(E) = h(p_e)$

(iii) $H(X|E,Y) = \underbrace{p(E=0)}_{=p_e} \underbrace{H(X|E=0,Y)}_{\leq H(X) \leq \log|\mathcal{X}|} + p(E=1) \underbrace{H(X|E=1,Y)}_{=0}$

$\square$

remark: if $\text{range}(Y) = \text{range}(X)$, we can replace $|\mathcal{X}|$ by $|\mathcal{X}-1|$.
In particular:

Corollary: If $\text{range}(Y) = \text{range}(X) = \{0,1\}$, then

$$\boxed{h(p_e) \geq H(X|Y)}$$

Corollary: Let $X = (X_1, \dots, X_n)$ describe a random n-bit string, Y a random variable, $\{f_i : \mathbb{R} \to \mathbb{R}\}_{i=1}^n$ and $p_e := \frac{1}{n} \sum_{i=1}^n p(X_i \neq f_i(Y))$ the "average bit-error rate". Then

$$\boxed{h(p_e) \geq \frac{1}{n} H(X|Y)}$$

proof: $\rightarrow$ exercise ...

II.7. Entropy rates

Def.: ∘ A "stochastic process" $\{X_i\}_{i\in\mathbb{N}}$ is a sequence of random variables.
It is called "stationary" iff $\forall n\in\mathbb{N}$

$p(X_{l+1}=x_1,\dots,X_{l+n}=x_n)$ is independent of $l\in\mathbb{N}_0$ for all x's.

∘ The "entropy rates" of a stochastic process are defined as

$$H(\{X_i\}) := \lim_{n\to\infty} \frac{1}{n} H(X_1,\dots,X_n),$$

$$H'(\{X_i\}) := \lim_{n\to\infty} H(X_n \mid X_{n-1},\dots,X_1)$$

if the limits exist.

Thm.: For a stationary stochastic process the entropy rates both exist and

$$\boxed{H(\{X_i\}) = H'(\{X_i\})}$$

proof:

$$H(X_{n+1}\mid X_n,\dots,X_1) \overset{\text{strong sub-additivity}}{\leq} H(X_{n+1}\mid X_n,\dots,X_2) \underset{\text{stationarity}}{=} H(X_n \mid X_{n-1},\dots,X_1)$$

→ H' exists since $H(X_n\mid X_{n-1},\dots,X_1)$ is a non-increasing and non-negative sequence.

The chain rule implies:

$$\lim_{n\to\infty}\frac{1}{n}H(X_n,\dots,X_1) = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} H(X_i\mid X_{i-1},\dots,X_1)$$

$$= \lim_{n\to\infty} H(X_n\mid X_{n-1},\dots,X_1)$$

Lemma (Cesaro mean): $a_n \to a \Rightarrow \frac{1}{n}\sum_{i=1}^{n} a_i \to a$

Corollary: For a stationary Markov chain $X_1 - X_2 - \dots$ we have

$$H(\{X_i\}) = H'(\{X_i\}) = H(X_2 | X_1)$$

proof: $H(X_n | X_{n-1}, \dots, X_1) \underset{\text{Markov}}{=} H(X_n | X_{n-1}) \underset{\text{stationary}}{=} H(X_2 | X_1)$ $\square$

Optimality of Huffman codes

Recursive construction of a „Huffman code" $C: \mathcal{X} \to A^+$ by building an $|A|$-ary tree from its leaves:

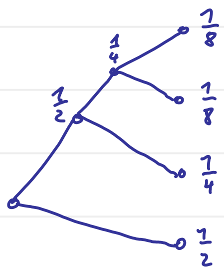
step 0) assign each symbol from $\mathcal{X}$ to a leaf,

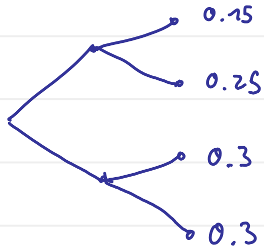
step 1) assign the $|A|$ least probable symbols to leaves with a common vertex,

step 2) 'combine' these symbols to a single one whose probability equals the sum of the $|A|$ probabilities.

Then iterate 1) → 2) → 1) → ... until there is only one symbol (= root) left.

Examples:





Thm.: Given a random variable X and an alphabet A. The Huffman code is optimal in the set of all prefix-free symbol codes $C: \mathcal{X} \to A^+$.

proof: by induction (only $|A|=2$):

- basis for induction: optimality holds for $n := |\mathcal{X}| = 2$
- induction Hypothesis: Huffman code is optimal for n. (*)
- induction step: assume (*), let $|\mathcal{X}| = n+1$ and $p(x_n), p(x_{n+1})$ be smallest probabilities. Then

$$\underbrace{L(C_{n+1}) = L(C_n)}_{\text{Huffman codes}} + p(x_n) + p(x_{n+1})$$

since $L(C_{n+1}) = l_1 p(x_1) + \dots l_{n-1} p(x_{n-1}) + l_n p(x_n) + l_{n+1} p(x_{n+1})$ with $l_n = l_{n+1}$

$L(C_n) = \quad$ —"— $\quad + (l_n - 1)(p(x_n) + p(x_{n+1}))$

Now assume $L(C'_{n+1}) < L(C_{n+1})$ for an optimal prefix-free code C'_{n+1}.

Optimality $\Rightarrow$ $l'_n = l'_{n+1} = \max\{l'_i\}$ and we can assume x_n, x_{n+1} to be neighbors on the tree of C'_{n+1}

Then $L(C_n) \overset{(*)}{\leq} L(C'_n) = L(C'_{n+1}) - p(x_{n+1}) - p(x_n)$

$< L(C_{n+1}) - p(x_{n+1}) - p(x_n) = L(C_n)$ ↯

$\square$

Exp.: Huffman code for English language (see MacKay)

$L(C) = 4.15$ bits, compared to $H(X) = 4.11$ bits

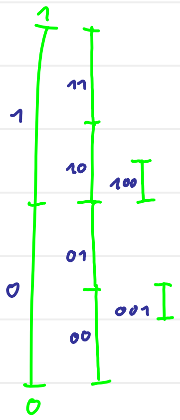
(entropy rate, however, is about 1bit/symbol)

remarks:
- Huffman codes are used in the final level of the JPEG algorithm,
- since any strategy for the Bar Kochba game corresponds to a prefix-free code and vice versa, Huffman codes provide the optimal strategy.

III.2. Stream codes

Motivation: "guessing game" (→ see MacKay)

III.2.1. Arithmetic codes



Basic idea: • $I : \{0,1\}^+ \to \{ [a,b) \mid 0 \le a < b \le 1 \}$

$$I(a_1 \dots a_n) := \left[\sum_{k=1}^{n} a_k 2^{-k} , \sum_{k=1}^{n} a_k 2^{-k} + 2^{-n} \right)$$

Note: $I(a_1 \dots a_n) \supseteq I(a_1 \dots a_{n+1})$

• Define $J : \mathcal{X}^+ \to \{ [a,b) \mid 0 \le a < b \le 1 \}$ similarly, but s.t. $|J(x_1 \dots x_n)| = p(x_1, \dots, x_n)$

• Encode $x_1 \dots x_n$ into $a_1 \dots a_k$ s.t.

$I(a_1 \dots a_k) \subseteq J(x_1 \dots x_n)$ & k is smallest possible.

• Encoding & decoding can be done 'on the fly'

• For a practicable algorithm J is constructed s.t. it depends only on a window of a fixed number of x_i's.

• arithmetic coding requires a model for the probabilities

• Applications: • Dasher

• DjVu

III.2.2. Lempel-Ziv coding

Idea: Replace a substring by a pointer to an earlier occurence of the same substring.

Example:

original string:	1	0	11	01	010	00	0101	01010
# of substring:	1	2	3	4	5	6	7	8
(pointer, additional bit):	(0,1)	(0,0)	(1,1)	(2,1)	(4,0)	(2,0)	(5,1)	(7,0)

remarks:

- applied in compress & gzip
- does not require a probabilistic model for the source
- Lempel-Ziv coding compresses asymptotically down to the entropy rate (for ergodic stationary stochastic processes)
- for too short strings the 'compressed' message can be longer than the original one
- a variant of LZ (Lempel-Ziv-Welch) is used in the image format GIF.

III.3. Non-perfect data compression

Thm.: If $C: \mathcal{X}^N \to \{0,1\}^+$ is a code for which $H(X^N) \geq L(C)$,

then
$$p_e \geq h^{-1}\left(\frac{H(X^N) - L(C)}{N}\right)$$

where h^{-1} is the inverse of the binary entropy function h on $[0, \frac{1}{2})$,

and p_e the average bit error rate after decoding.

proof: define a random variable $Y := C(X^N)$ with range $\mathcal{Y}$ and a code $C': \mathcal{Y} \to \{0,1\}^+$ by $C' = id$. Then

$$L(C) = L(C') \geq H(Y) = H(C(X^N)) \quad (*)$$

$$H(X^N | C(X^N)) = H(X^N, C(X^N)) - H(C(X^N))$$

$$\overset{H(C(X^N)|X^N)=0}{=} H(X^N) - H(C(X^N))$$

$$\overset{(*)}{\geq} H(X^N) - L(C)$$

Fano's inequality $\Rightarrow$
$$N h(p_e) \geq H(X^N | C(X^N)) \geq H(X^N) - L(C)$$

$\square$

III.4. Asymptotic equipartition property & typicality

Def.: A sequence of random variables $\{X_i\}_{i\in\mathbb{N}}$ converges to X "in probability" if $\forall \varepsilon > 0$ $p\{|X_n - X| > \varepsilon\} \to 0$ for $n \to \infty$.

Thm.: (weak law of large numbers)

Let $\{X_i\}_{i\in\mathbb{N}}$ be i.i.d. random variables with mean $E(X_i) = \mu$.

Then $\boxed{\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{n\to\infty} \mu}$ in probability.

Thm.: (asymptotic equipartition property (AEP)) Let $\{X_i\}_{i\in\mathbb{N}}$ be i.i.d. random variables with distribution $p(x)$. Then

$$-\frac{1}{n}\log\big(p(X_1,\dots,X_n)\big) \to H(X) \quad \text{in probability.}$$

remark: $p(X)$ means a random variable defined as follows: let Ω be the sample space, $X:\Omega\to\mathcal{X}$ a r.v. and $p:\mathcal{X}\to[0,1]$, $x\mapsto p(X=x)$. Then $p(X):\Omega\to[0,1]$ is defined as $p(X)=p\circ X$. Hence if μ is the probability measure on Ω, then $p(X):\omega\mapsto\mu\big(\{\omega'\in\Omega \mid X(\omega)=X(\omega')\}\big)$.

proof: $\{X_i\}$ i.i.d. $\Rightarrow$ $\{\log p(X_i)\}$ i.i.d.

Law of large numbers $\Rightarrow$ $\frac{1}{n}\sum_{i=1}^{n}\log(p(X_i)) \to -H(X)$

$$\overset{\shortparallel}{\frac{1}{n}\log\big(p(X_1,\dots,X_n)\big)}$$

remark: this can be extended to ergodic stationary stochastic processes. The "Shannon-McMillan-Breiman thm." states that then

$$-\frac{1}{n}\log\big(p(X_1,\dots,X_n)\big) \to H(\{X_i\})$$

Def.: (typical set)

A "typical set" $A_\varepsilon^{(n)}\subseteq\mathcal{X}^n$ w.r.t. to a set of i.i.d. random variables $\{X_i\}_{i\in\mathbb{N}}$ contains all $x\in\mathcal{X}^n$ for which

$$2^{-n(H(X)+\varepsilon)} \leq p(x) \leq 2^{-n(H(X)-\varepsilon)}$$

Motivation: take a random string $x := (x_1,\dots,x_n)\in\{1,\dots,k\}^n$

Then $p(x) = \prod_{j=1}^{k} p(X=j)^{n_j} \underset{\uparrow}{\approx} 2^{-nH(X)}$

$n_j \approx n\,p(X=j)$

$\to$ we expect a random string to have probability around $2^{-nH(X)}$

Thm.: (properties of sets of typical strings)

1) $x \in A_\varepsilon^{(n)} \iff H(X) - \varepsilon \leq -\frac{1}{n} \log p(x) \leq H(X) + \varepsilon$

2) $p(A_\varepsilon^{(n)}) := p\{X^n \in A_\varepsilon^{(n)}\} > 1 - \varepsilon$ for suff. large n

3) $|A_\varepsilon^{(n)}| \leq 2^{n(H(X)+\varepsilon)}$

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

proof: 1) from definition

2) $p(A_\varepsilon^{(n)}) = p\{|-\frac{1}{n}\log p(X) - H(X)| \leq \varepsilon\}$

$$= 1 - p\{|-\tfrac{1}{n}\log p(X) - H(X)| > \varepsilon\}$$

AEP $\Rightarrow \exists N \in \mathbb{N}\ \forall n > N: p\{|-\frac{1}{n}\log p(X) - H(X)| > \varepsilon\} < \varepsilon$

3) $$1 \geq \sum_{x \in A_\varepsilon^{(n)}} p(x) \geq \sum_{x \in A_\varepsilon^{(n)}} 2^{-n(H(X)+\varepsilon)} = |A_\varepsilon^{(n)}| \, 2^{-n(H(X)+\varepsilon)}$$

4) from 2) $\Rightarrow \exists N \in \mathbb{N}\ \forall n > N: p(A_\varepsilon^{(n)}) > 1 - \varepsilon$

$$\wedge\!\!\!\!\wedge$$

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

$\square$

loosely speaking:
- sequences are typically typical ones
- there are $\sim 2^{nH(X)}$ typical sequences of length n
- each of them occurs with probability $\sim 2^{-nH(X)}$

III.5. Data compression based on AEP

$\mathcal{X}^n = A_\varepsilon^{(n)} \cup \overline{A_\varepsilon^{(n)}}$

Define an injective map $C: \mathcal{X}^n \to \{0,1\}^+$ such that

- $x \in A_\varepsilon^{(n)} \Rightarrow C(x) = 0y$ where $y \in \{0,1\}^{\lceil n(H(X)+\varepsilon)\rceil}$
  - remember that $|A_\varepsilon^{(n)}| \leq 2^{n(H(X)+\varepsilon)}$
    $\to$ C can be chosen injective on $A_\varepsilon^{(n)}$
  - the prefix "0" indicates that $x \in A_\varepsilon^{(n)}$
- $x \notin A_\varepsilon^{(n)} \Rightarrow C(x) = 1\tilde{x}$ where $\tilde{x} \in \{0,1\}^{\lceil n \log|\mathcal{X}|\rceil}$
  - $\tilde{x} = x$ if $\mathcal{X} = \{0,1\}$
  - the prefix "1" encodes $x \notin A_\varepsilon^{(n)}$

We obtain for the average codeword length:

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

$$\leq \underbrace{p(A_\varepsilon^{(n)})}_{\leq 1}\big(n(H(X)+\varepsilon)+2\big) + \underbrace{\big(1-p(A_\varepsilon^{(n)})\big)}_{\leq \varepsilon}\big(n\log|\mathcal{X}|+2\big)$$

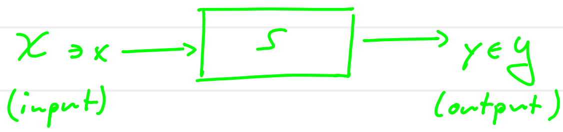
$$\leq n(H(X)+\varepsilon) + 2 + \varepsilon\big(n\log|\mathcal{X}|+2\big)$$

$\Rightarrow$ Thm.: (Shannon's source coding theorem) Let $\{X_i\}_{i\in\mathbb{N}}$ be i.i.d. random variables with range $\mathcal{X}$. $\forall \delta > 0\ \exists n \in \mathbb{N}\ \exists C: \mathcal{X}^n \to \{0,1\}^+$ uniquely decodable:

$$\boxed{\frac{1}{n}L(C) \leq H(X) + \delta}$$

IV. Shannon's noisy channel coding theorem

IV.1. Discrete memoryless channels



Def.: Let $\mathcal{X}$ and $\mathcal{Y}$ be finite sets. A map $S: \mathbb{R}_+^{|\mathcal{X}|} \to \mathbb{R}_+^{|\mathcal{Y}|}$ describes a "discrete memoryless channel" and the characterizing matrix $S \in \mathbb{R}_+^{|\mathcal{Y}| \times |\mathcal{X}|}$ is a "stochastic matrix" if $S_{yx} =: p(y|x)$ are conditional probabilities, i.e., $\forall x \in \mathcal{X}: \sum_{y \in \mathcal{Y}} p(y|x) = 1$.

remarks:

- "memoryless" refers to the fact that n uses of the channel will be described by $S^{\otimes n} := S \otimes \dots \otimes S : \mathbb{R}_+^{|\mathcal{X}^n|} \to \mathbb{R}_+^{|\mathcal{Y}^n|}$

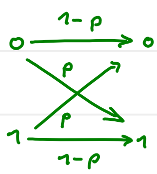
  where $\left(S^{\otimes n}\right)_{y,x} = \prod_{i=1}^{n} p(y_i | x_i)$, $x \in \mathcal{X}^n$, $y \in \mathcal{Y}^n$

  That is, the transition probabilities do not depend on what was sent in the past.

- in the following "channel" is meant to be discrete & memoryless

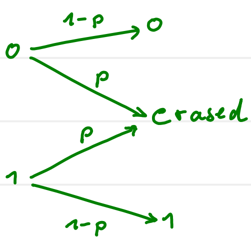
Examples:

- binary symmetric channel: $S = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}$, $p \in [0,1]$



"bit flip" occurs with prob. p

- binary erasure channel: $S = \begin{pmatrix} 1-p & 0 \\ p & p \\ 0 & 1-p \end{pmatrix}$



bit is lost with prob. p

IV.2. Codes, errors and rates

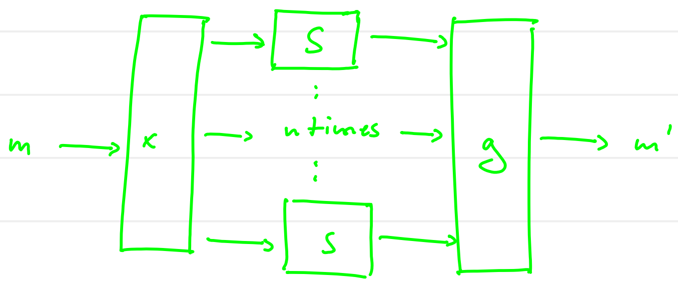
Def.: An "(M,n) code" for a channel S with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$ consists of

(i) an index set $\mathcal{M}$ (= set of messages) with $|\mathcal{M}| = M$

(ii) an encoding function $x : \mathcal{M} \to \mathcal{X}^n$
with "codewords" $x(m), m \in \mathcal{M}$ and "codebook" $x(\mathcal{M})$

(iii) a decoding function $g : \mathcal{Y}^n \to \mathcal{M}$

n is called "blocklength"



Example: repetition code: $x : \{0,1\} \to \{0,1\}^3$, $0 \mapsto 000$, $1 \mapsto 111$
$g : \{0,1\}^3 \longrightarrow \{0,1\}$ by majority vote

Errors:

- conditional prob. of error: $\lambda_m := \sum_{y : g(y) \neq m} p(y \mid x(m))$, $m \in \mathcal{M}$
  where $p(y \mid x(m)) = \prod_{i=1}^{n} p(y_i \mid x_i(m))$
- max. prob. of error: $\lambda^{(n)} := \max \{ \lambda_m \}_{m \in \mathcal{M}}$
- average prob. of error: $p_e^{(n)} := \frac{1}{M} \sum_{m \in \mathcal{M}} \lambda_m$

Def.: • The "rate" of an (M,n) code is $R := \frac{\log M}{n}$ (bits/transmission)

• A rate R is called "achievable" for a given channel iff there exists a sequence of $(\lceil 2^{nR} \rceil, n)$ codes s.t. $\lambda^{(n)} \to 0$ as $n \to \infty$.

• The "capacity" $C(S)$ of a channel S is the supremum over all achievable rates.

remark: the rate of a repetition code $0 \mapsto 0^n$, $1 \mapsto 1^n$ is $R = \frac{1}{n}$.
So if we require $\lambda^{(n)} \to 0$, then $R \to 0$ for generic channels.

IV.3. Joint AEP

Def.: Let $n \in \mathbb{N}$, $\varepsilon > 0$ and $p : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be the joint distribution of random variables X and Y with ranges $\mathcal{X}, \mathcal{Y}$. The set of jointly typical sequences w.r.t. to the joint distribution p is defined as

$$B_\varepsilon^{(n)} := \Big\{ (x,y) \in \mathcal{X}^n \times \mathcal{Y}^n \;\Big|\; \left| -\tfrac{1}{n} \log p(x) - H(X) \right| < \varepsilon \;\wedge\; \left| -\tfrac{1}{n} \log p(y) - H(Y) \right| < \varepsilon \;\wedge\; \left| -\tfrac{1}{n} \log p(x,y) - H(X,Y) \right| < \varepsilon \Big\}$$

where $p(x,y) := \prod_{i=1}^{n} p(x_i, y_i)$ and $p(x)$ & $p(y)$ are the marginals.

Thm.: (joint AEP) Let $B_\varepsilon^{(n)}$ be the set of jointly typical sequences w.r.t. the joint distribution of X and Y. Then

1) $p(B_\varepsilon^{(n)}) > 1-\varepsilon$ for n suff. large

2) $|B_\varepsilon^{(n)}| \leq 2^{n(H(X,Y)+\varepsilon)}$ $\forall n \in \mathbb{N}$

3) $|B_\varepsilon^{(n)}| \geq (1-\varepsilon)\, 2^{n(H(X,Y)-\varepsilon)}$ for n suff. large

4) If $\tilde{X}^n, \tilde{Y}^n$ are i.i.d. random variables with individual ranges $\mathcal{X}$ and $\mathcal{Y}$ and joint distribution $Pr(\tilde{X}=x, \tilde{Y}=y) =: \tilde{p}(x,y)$ of the form $\tilde{p}(x,y) = p(x)p(y)$ where $p(x)$ and $p(y)$ are the marginal distributions of X and Y respectively. Then

a) $(1-\varepsilon)\, 2^{-n(I(X;Y)+3\varepsilon)} \leq Pr((\tilde{X}^n, \tilde{Y}^n) \in B_\varepsilon^{(n)})$ for n suff. large,

b) $Pr((\tilde{X}^n, \tilde{Y}^n) \in B_\varepsilon^{(n)}) \leq 2^{-n(I(X;Y)-3\varepsilon)}$

proof: 1), 2) and 3) are proven in complete analogy to the AEP for $A_\varepsilon^{(n)}$.

4) a) $$Pr((\tilde{X}^n, \tilde{Y}^n) \in B_\varepsilon^{(n)}) = \sum_{(x,y)\in B_\varepsilon^{(n)}} p(x)p(y)$$

$$\leq |B_\varepsilon^{(n)}|\, 2^{-n(H(X)-\varepsilon)}\, 2^{-n(H(Y)-\varepsilon)}$$

$$\overset{2)}{\leq} 2^{-n(I(X;Y)-3\varepsilon)}$$

b) $$Pr((\tilde{X}^n, \tilde{Y}^n) \in B_\varepsilon^{(n)}) = \sum_{(x,y)\in B_\varepsilon^{(n)}} p(x)p(y)$$

$$\geq |B_\varepsilon^{(n)}|\, 2^{-n(H(X)+\varepsilon)}\, 2^{-n(H(Y)+\varepsilon)}$$

$$\overset{3)}{\geq} (1-\varepsilon)\, 2^{-n(I(X;Y)+3\varepsilon)}$$

$\square$

IV.4. Direct part of the coding theorem

<u>Thm.:</u> Let $p(y|x)$ with $x \in \mathcal{X}, y \in \mathcal{Y}$ describe a discrete memoryless channel. Every $R < \max_{p(x)} I(X;Y)$ is an achievable rate for it, if the mutual information is computed w.r.t. to $p(x,y) := p(x)\,p(y|x)$.

<u>proof:</u> • fix $\varepsilon > 0$, $n \in \mathbb{N}$, $p(x)$ and let $\mathcal{M} := \{1, \dots, 2^{nR}\}$

• produce "random" $(2^{nR}, n)$ code by generating 2^{nR} codewords in $\mathcal{X}^n$ independently according to $p(x^n) = \prod_{i=1}^{n} p(x_i)$

• use "typical-set decoding" $g: \mathcal{Y}^n \to \mathcal{M}$ defined by

$$g(y^n) = m \quad \text{if} \quad (x^n(m), y^n) \in B_\varepsilon^{(n)}$$
$$\wedge \; \forall_{j \neq m}: \; (x^n(j), y^n) \notin B_\varepsilon^{(n)}$$

$$g(y^n) = 1 \quad \text{otherwise}.$$

<u>error analysis:</u> $\hat{p} := \sum_c p(c)\, p_e^{(n)}(c)$ averaged over codes

$$= \sum_c p(c)\, 2^{-nR} \sum_{m \in \mathcal{M}} \lambda_m(c) \qquad \text{\& codewords}$$

$$= 2^{-nR} \sum_m \underbrace{\sum_c p(c)\, \lambda_m(c)}_{\text{independent of } m} = \sum_c p(c)\, \lambda_1(c)$$

two error types (assuming y^n is received upon sending $x^n(1)$):

(i) $(x^n(1), y^n) \notin B_\varepsilon^{(n)}$; this has prob. at most ε

(ii) $(x^n(j), y^n) \in B_\varepsilon^{(n)}$ for some $j \neq 1$

Since $X^n(1)$ and $X^n(j)$ are independent if $j \neq 1$, so are Y^n & $X^n(j)$

$\overset{\text{joint AEP}}{\Longrightarrow}$ prob. bounded by $2^{-n(I(X:Y)-3\varepsilon)}$ for each $j \neq 1$

$$\Rightarrow \quad \hat{p} \leq \overset{(*)}{\varepsilon} + (2^{nR}-1)\, \overbrace{2^{-n(I(X:Y)-3\varepsilon)}}^{(**)}$$

$$\leq \varepsilon + 2^{n(R - I(X:Y) + 3\varepsilon)}$$

$$\leq 2\varepsilon \quad \text{if} \quad R < I(X:Y) - 3\varepsilon \;\; \& \;\; n \text{ suff. large}$$

Hence, if $R < I(X:Y)$ we can choose $\varepsilon > 0$ and $n \in \mathbb{N}$ accordingly and make $\hat{p}$ arbitrary small.

- $\hat{p} \leq 2\varepsilon \Rightarrow \exists c: \; P_e^{(n)}(c) \leq 2\varepsilon$
- modify this code by discarding the worst 50% codewords

  $\to (2^{nR-1}, n)$ code $\tilde{c}$ for which the max. prob. of error is

$$\lambda^{(n)}(\tilde{c}) \leq 2 P_e^{(n)}(c) \leq 4\varepsilon$$

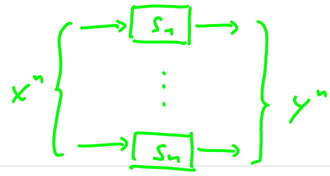
The rate of $\tilde{c}$ is $\tilde{R} = R - \frac{1}{n}$ $\square$

IV.5. Converse part of the coding theorem

Lemma: Let $x^n \in \mathcal{X}^n$ with distribution $p(x^n)$ be the input and $y^n \in \mathcal{Y}^n$ be the output of an n-fold product of discrete memoryless channels. For $p(x^n, y^n) := \prod_{j=1}^{n} p_j(y_j | x_j)\, p(x^n)$ we get

$$\boxed{I(X^n : Y^n) \leq \sum_{j=1}^{n} I(X_j : Y_j)}$$

(note that the channels can be different)



<u>proof:</u> $I(X^n:Y^n) = H(Y^n) - H(Y^n|X^n)$

chain rule

$$\stackrel{\downarrow}{=} H(Y^n) - \sum_{j=1}^{n} H(Y_j|Y_{j-1},\dots,Y_1,X^n)$$

Y_j only depends on X_j

$$\stackrel{\downarrow}{=} H(Y^n) - \sum_{j=1}^{n} H(Y_j|X_j)$$

subadditivity

$$\stackrel{\downarrow}{\leq} \sum_{j=1}^{n} H(Y_j) - H(Y_j|X_j) = \sum_{j=1}^{n} I(X_j:Y_j)$$

$\square$

<u>Thm.:</u> (Shannon's noisy coding theorem - converse part)

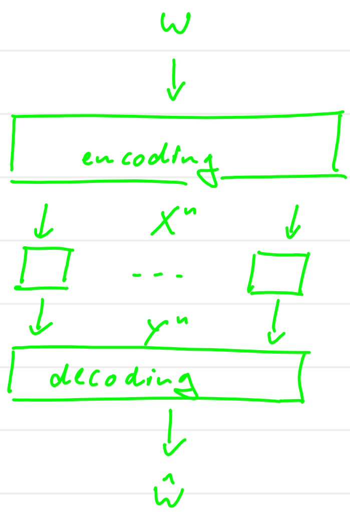
Any $(2^{nR}, n)$ code for a discrete memoryless channel satisfies

$$\boxed{R \leq \frac{C}{1-p_e^{(n)}}} \quad \text{where} \quad \boxed{C := \max_{p(x)} I(X:Y)} \quad \text{and}$$

$p_e^{(n)} := 2^{-nR} \sum_{m=1}^{2^{nR}} \lambda_m$ is the error prob. averaged over all codewords.

<u>proof:</u> Let W be a random variable assigned to uniform dist. of codewords.

That is, $\text{range}(W) = \{1, \dots, 2^{nR}\}$



$$nR = H(W) = H(W|\hat{W}) + I(W:\hat{W})$$

Fano's inequality $\rightarrow$ $\leq h(p_e^{(n)}) + p_e^{(n)} nR + I(W:\hat{W})$

data processing ineq. for Markov chain $W - X^n - Y^n - \hat{W}$ $\leq h(p_e^{(n)}) + p_e^{(n)} nR + I(X^n:Y^n)$

previous Lemma $\longrightarrow$ $\leq h(p_e^{(n)}) + p_e^{(n)} nR + \sum_{j=1}^{n} I(X_j:Y_j)$

$$\leq h(p_e^{(n)}) + p_e^{(n)} nR + nC$$

$$\Rightarrow \quad p_e^{(n)} \geq 1 - \frac{C}{R} - \frac{h(p_e^{(n)})}{nR}$$

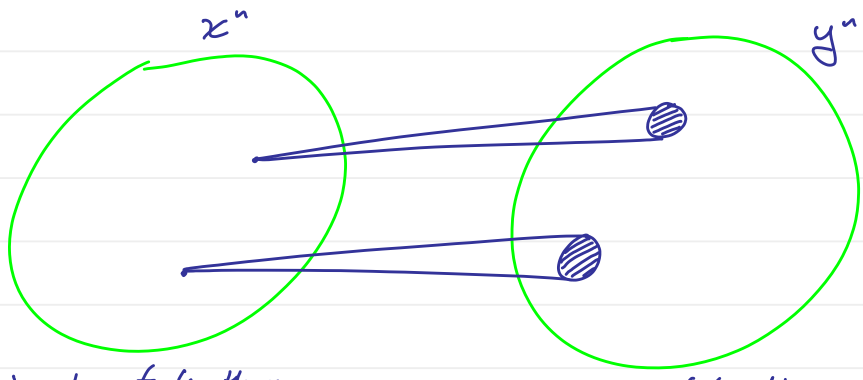
$\forall m \in \mathbb{N}$ we can construct a $(2^{nmR}, nm)$ code for which $p_e^{(nm)} = p_e^{(n)}$.

$$\Rightarrow \quad p_e^{(n)} = p_e^{(nm)} \geq 1 - \frac{C}{R} - \frac{h(p_e^{(nm)})}{nmR} \xrightarrow{m \to \infty} 1 - \frac{C}{R}$$

$$\Rightarrow \quad R \leq \frac{C}{1 - p_e^{(n)}}$$

$\square$

IV.6. Heuristic view on Shannon's noisy channel coding theorem



(i) for every input we obtain $\sim 2^{H(Y|X)}$ outputs with roughly equal prob. The other outputs may be possible, but they are not typical and thus ε unlikely.

(ii) the total number of typical sequences at the output is $\sim 2^{nH(Y)}$

(iii) to be able to distinguish different inputs at the output, there images must not overlap.

$\rightarrow$ there are about $\dfrac{2^{nH(Y)}}{2^{nH(Y|X)}} = 2^{nI(X:Y)}$ such inputs corresponding to a rate of $\underline{\underline{I(X:Y)}}$

<u>remark</u>: strictly speaking, in (i) it should be $2^{nH(Y|X=x)}$ for an input $x \in \mathcal{X}^n$

IV.7. Properties of the channel capacity

Proposition: (1) $C \geq 0$

(2) $C \leq \min\{\log|\mathcal{X}|, \log|\mathcal{Y}|\}$

proof: both are rather obvious from the operational definition, but they also follow easily from properties of the mutual information:

(1) $I(X:Y) \geq 0$

(2) $$I(X:Y) = \begin{cases} H(X) - H(X|Y) \leq H(X) \leq \log|\mathcal{X}| \\ H(Y) - H(Y|X) \leq H(Y) \leq \log|\mathcal{Y}| \end{cases}$$ $\square$

Thm.: (additivity) Let $S_1, S_2, S_1 \otimes S_2$ be stochastic matrices describing two discrete memoryless channels and their product, respectively. Then

$$\boxed{C(S_1 \otimes S_2) = C(S_1) + C(S_2)}$$

proof: $C(S_1 \otimes S_2) = \max\limits_{p(x^2)} I(X^2 : Y^2)$ where $X^2 := (X_1, X_2)$, $x^2 \in \mathcal{X} \times \mathcal{X}$

(i) $\geq \max\limits_{p(x_1)p(x_2)} I(X^2 : Y^2) = C(S_1) + C(S_2)$

(ii) $\leq \max\limits_{p(x^2)} \big(I(X_1 : Y_1) + I(X_2 : Y_2) \big)$ due to additivity Lemma

$= \max\limits_{p(x_1)p(x_2)} \big(\quad " \quad + \quad " \quad \big) = C(S_1) + C(S_2)$

$\square$

Proposition: $I(X:Y)$ w.r.t. $p(x,y) = \overbrace{p(y|x)}^{=:S_{yx}} p(x)$ is a convex functional of $p(y|x)$ and a concave functional of $p(x)$.

proof:
- $I(X:Y) = \underbrace{H(Y)}_{\text{concave in } p(x) \text{ since } p(y) = \sum_x S_{yx} p(x) \text{ depends linearly on } p(x) \text{ \& } H \text{ is concave}} - H(Y|X) = H(Y) - \underbrace{\sum_x p(x) \overbrace{H(Y|X=x)}^{\text{depends only on } S \text{ not on } p(x)}}_{\text{linear in } p(x)}$

  $\Rightarrow$ concavity in $p(x)$

- for the proof of convexity w.r.t. S fix $p(x)$ and define

  $p_\lambda(y|x) := \lambda p_1(y|x) + (1-\lambda) p_2(y|x)$, $p_\lambda(x,y) := p_\lambda(y|x) p(x)$

  Then $I_\lambda(X:Y) = D(p_\lambda(x,y) \| p_\lambda(y) p(x))$ and convexity follows from joint convexity of the relative entropy. $\square$

Corollary:
- The channel capacity is a convex functional of the channel:

$$C(\lambda S_1 + (1-\lambda) S_2) \leq \lambda C(S_1) + (1-\lambda) C(S_2)$$

- For $\max_{p(x)} I(X:Y)$ any local maximum is a global one.

$\rightarrow$ efficient algorithms for computing the capacity (e.g. Arimoto-Blahut)

IV.8. Computing some capacities

Prop.: Let S with $S_{yx} = p(y|x)$ be a stochastic matrix where all columns are permutations of a probability vector q. Then

$$C(S) = \left(\max_{p(x)} H(Y) \right) - H(q) \qquad \left(\text{where } Y \text{ is distributed according to } \sum_x p(y|x)p(x) \text{ and } y \in \mathcal{Y} \right)$$

$$\leq \log |\mathcal{Y}| - H(q)$$

with equality iff there is an input distribution $\tilde{p}$ s.t. $(S\tilde{p})_y = \frac{1}{|\mathcal{Y}|}$ $\forall y \in \mathcal{Y}$.

proof: $C = \max_{p(x)} I(X;Y) = \sup_{p(x)} H(Y) - H(Y|X)$

$$= \sup_{p(x)} \underbrace{H(Y)}_{\leq \log |\mathcal{Y}|} - \sum_x p(x) \underbrace{H(Y|X=x)}_{= H(q)}$$

and $H(Y) = \log |\mathcal{Y}|$ iff distribution is uniform. $\square$

Examples: ① binary symmetric channel:

$$S = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix} \qquad (\text{bit flipped with prob. } p)$$

$$\boxed{C(S) = \log 2 - h(p)} \qquad \text{for } \tilde{p} = \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}$$

② noisy typewriter channel:

$$S = \begin{pmatrix} 1-2p & p & & & p \\ p & 1-2p & p & & \\ & p & \ddots & \ddots & \\ & & \ddots & \ddots & p \\ p & & & p & 1-2p \end{pmatrix} \qquad \text{"circulant matrix"}$$

$$\boxed{C(S) = \log |\mathcal{Y}| - H(q)}, \quad q = (1-2p, p, p) \qquad \text{for } \tilde{p} \text{ uniform}$$

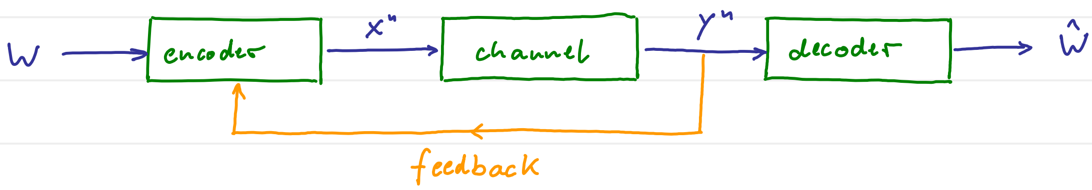
③ binary erasure channel

$$S = \begin{pmatrix} 1-p & 0 \\ p & p \\ 0 & 1-p \end{pmatrix}$$ (: bit erased with prob. p)

$$\boxed{C(S) = 1-p}$$

(proven in the exercise. A uniform output distribution is in this case not possible. Uniform input $\tilde{p}$ is optimal, though.)

IV.9. Feedback capacity



Consider sequential uses of the channel where in each encoding step the output of all previous transmissions can be used.

Def.: • An (M,n)-code with feedback for a discrete memoryless channel with input & output alphabets $\mathcal{X}$ & $\mathcal{Y}$ is defined via

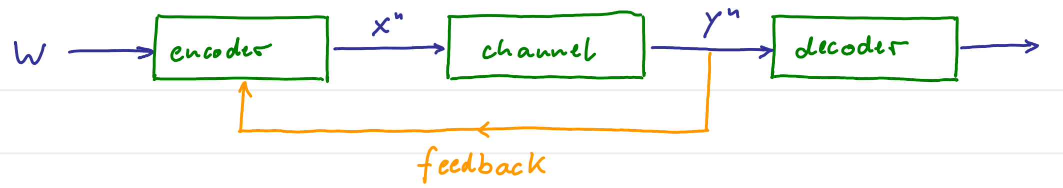
- encoding functions $f_i : \{1,\dots,M\} \times \mathcal{Y}^{i-1} \to \mathcal{X}$, $i=1,\dots,n$
- a decoding function $g : \mathcal{Y}^n \to \{1,\dots,M\}$

• Let $Y^i := (Y_1,\dots,Y_i)$ and $X_i := f_i(W, Y^{i-1})$

• R is a "rate achievable with feedback" iff $\forall \varepsilon > 0\ \exists\ (2^{nR}, n)$-code with feedback s.t. $\lambda^{(n)} < \varepsilon$.

• The "feedback capacity" C_{FB} is the supremum over all such rates.

Thm.: $$\boxed{C_{FB} = C}$$



Thm.: $\boxed{C_{FB} = C}$

proof: evidently $C_{FB} \geq C$, so we need to show $C_{FB} \leq C$.

Let W be uniformly distributed over all input messages.

Similar to the proof of the converse part without feedback:

$$nR = H(W) = \underbrace{H(W|Y^n)}_{(i)} + \underbrace{I(W;Y^n)}_{(ii)}$$

(i) $H(W|Y^n) \leq H(W|\hat{W})$ data processing inequality

$\leq 1 + p_e^{(n)} nR$ Fano's inequality $(h(p_e) + p_e \log|\mathcal{X}| \geq H(X|Y))$

(ii) $I(W;Y^n) = H(Y^n) - H(Y^n|W)$

$= H(Y^n) - \sum_{i=1}^n H(Y_i|Y^{i-1}, W)$

$= H(Y^n) - \sum_{i=1}^n H(Y_i|Y^{i-1}, W, X_i)$ | $X_i = f_i(W, Y^{i-1})$

$= H(Y^n) - \sum_{i=1}^n H(Y_i|X_i)$ | Y_i depends on (Y^{i-1}, W) only via X_i

$\leq \sum_{i=1}^n H(Y_i) - H(Y_i|X_i)$ | subadditivity

$= \sum_{i=1}^n I(X_i;Y_i) \leq nC$

$\Rightarrow nR \leq 1 + p_e^{(n)} nR + nC \Rightarrow R(1 - p_e^{(n)}) \leq \frac{1}{n} + C$

$\Rightarrow R \leq C$ via $n \to \infty$

$\square$

remarks: in practice, however, feedback can help/simplify.

example: for the binary erasure channel resend the bit until it has not been erased $\longrightarrow$ average nr. of channels used:

$$(1-p)\underbrace{\sum_{n=1}^{\infty} n p^{n-1}}_{(1-p)^{-2}} = \frac{1}{(1-p)}$$

$\rightarrow$ $(1-p)$ is achievable rate with feedback.

But we know also that $C_{FB} = C = (1-p)$

note: • the capacity is in this case easily achieved with zero error

• without feedback codes coming close to capacity are far more complicated & the error is non-zero

• another resource which doesn't change capacity is "shared randomness" between sender & receiver.

IV. 10. Source-channel separation

Question: what if the messages to be transmitted are not uniformly distributed?

- one possibility is to separate source coding (data compression) & channel coding
- a more general approach would be to combine them. Such codes are called source-channel codes.
- the following shows that we don't loose anything, if we separate the two:

Thm.: (source-channel coding theorem)

Consider a discrete memoryless channel with $C := \max_{p(x)} I(X;Y)$.

(i) Let $\{V_i\}_{i\in\mathbb{N}}$ be a stochastic process which satisfies the AEP w.r.t. its entropy rate $H(\{V_i\})$ (e.g. an i.i.d source or, more generally, a stationary ergodic stochastic process). If $H(\{V_i\}) < C$, there is a source-channel code which allows transmission s.t. $\text{prob}(\hat{V}^n \neq V^n) \to 0$ as $n \to \infty$.

(ii) For any stationary stochastic process, if $H(\{V_i\}) > C$, then

$$\exists \delta > 0 \ \forall n \in \mathbb{N} \ \forall \text{ source-channel codes}: \ \text{prob}(\hat{V}^n \neq V^n) > \delta.$$

proof: similar to what we did before → exercise.

V. Error correcting codes / coding theory

Note: "random coding" (as in the proof of Shannon's noisy channel coding thm.) is completely useless for actual information transmission. We need something more concrete & more efficient ...

Example: "[7,4] Hamming code"

Let $x \in \{0,1\}^4$ be a message which we want to protect against errors.

Define $g := \begin{pmatrix} 0&1&1&1 \\ 1&0&1&1 \\ 1&1&0&1 \end{pmatrix}$ and $G := \begin{pmatrix} \mathbb{1}_4 \\ g \end{pmatrix} \in \mathbb{Z}_2^{7\times 4}$.

"Encode" the message into $y := Gx \in \mathbb{Z}_2^7$ with addition mod 2.

Claim: the image of any vector $x' \in \mathbb{Z}_2^4$ with $x' \neq x$ differs from y in at least three bits, i.e. $|\{i \mid (G(x-x'))_i \neq 0\}| \geq 3$.

Consequence: if an arbitrary single bit in y is corrupted, we can correct for it.

proven by inspection: $G(\Delta x)$ has at least 3 non-zero components if $\Delta x \neq 0$.

V.1. Basic definitions

Def.: Let $\mathcal{X}$ be a finite alphabet, $0 \in \mathcal{X}$ and $x, x' \in \mathcal{X}^n$.

- $d(x) := \left|\{ i \in \{1, \dots, n\} \mid x_i \neq 0 \}\right|$ "Hamming weight"
- $d(x-x') := |\{ i \mid x_i \neq x_i' \}|$ "Hamming distance"
- $\{ x' \in \mathcal{X}^n \mid d(x-x') \leq r \}$ "Hamming ball" of radius r around x

remark: $(x, x') \mapsto d(x-x')$ is a metric on $\mathcal{X}^n$.

Def.: An "error correcting code" C of length $n \in \mathbb{N}$ over an alphabet $\mathcal{X}$ is a subset $C \subseteq \mathcal{X}^n$ whose elements are called "codewords".

remarks:
- we will often associate an "encoding map" $E : \{1, \dots, |C|\} \to C \subseteq \mathcal{X}^n$ with the error correcting code (= code in the following)
- the above codes are also called "block codes" with "block length" n,
- the code is called q-ary (binary) if $|\mathcal{X}| = q$ $(|\mathcal{X}| = 2)$

Def.: Let $C \subseteq \mathcal{X}^n$.

- $R(C) := \dfrac{\log |C|}{\log |\mathcal{X}^n|}$ is called the "rate" of the code.
- $d(C) := \min\limits_{\substack{c, c' \in C \\ c \neq c'}} d(c-c')$ is called its "distance", and $\frac{d(C)}{n}$ "relative distance".

remarks:
- $R(C) \sim$ fraction of non-redundant info in the codewords of C.
- the [7,4] Hamming code has distance 3 & rate $R(C) = \frac{4}{7}$
- a corrupted message x' is said to have k errors w.r.t. its uncorrupted version x if $d(x-x') = k$.

Note: A code with distance d allows to correct

(i) $\lfloor \frac{(d-1)}{2} \rfloor$ errors,

(ii) $(d-1)$ symbol erasures.

proof: just choose the codeword closest in Hamming distance. $\square$

Def.: Let $\mathcal{C} := \{C_i\}_{i\in\mathbb{N}}$ be a sequence of codes with lengths n_i so that $n_{i+1} > n_i$. $\mathcal{C}$ is called "asymptotically good" if $\liminf_i R(C_i)$ and $\liminf_i \left(\frac{d(C_i)}{n_i}\right)$ are both strictly positive.

Summary of basic notions from previous lecture:

- "error correcting code" $\quad C \subseteq \mathcal{X}^n$ with $\begin{cases} \mathcal{X}: & \text{finite alphabet} \\ n \in \mathbb{N}: & \text{"length" of the code} \end{cases}$

- "rate" of an ECC, $\quad R(C) := \dfrac{\log |C|}{n \log |\mathcal{X}|} \sim \dfrac{\text{length of message}}{\text{length of its codeword}}$

  $\sim$ fraction of non-redundant info

- "distance" of an ECC: $\quad d(C) := \min\limits_{\substack{c, c' \in C \\ c \neq c'}} d(c - c')$ = min. Hamming distance between two codewords

- "relative distance": $\quad \dfrac{d(C)}{n}$

remember: an ECC with $d := d(C)$ allows to correct $\lfloor \frac{d-1}{2} \rfloor$ errors or $(d-1)$ symbol erasures

V.2. Linear codes

Def.: If $\mathcal{X}$ is a field and $C \subseteq \mathcal{X}^n$ a subspace, then C is calle a "linear code".

remarks:
- $|\mathcal{X}| < \infty$ implies that $\mathcal{X} = GF(q)$ is a "Galois field" with $q := |\mathcal{X}| = p^m$ for some prime p and $m \in \mathbb{N}$.
- A subspace $C \subseteq GF(q)^n$ admits a basis $c_1, \dots, c_k$ so that

  $\boxed{|C| = q^k}$ & thus $\boxed{R(C) = \frac{k}{n}}$

- for real world applications we often have $n \sim 10^3 - 10^4$

Def.:
- $G \in GF(q)^{n \times k}$ is called a "generator matrix" for a linear code $C \subseteq GF(q)^n$ if its columns form a basis of C.
- C is then called an "$[n,k]$-code" or "$[n,k,d]$-code" if $d = d(C)$.

remark: the encoding map $E: GF(q)^k \to GF(q)^n$ of a linear code is then just $E: x \mapsto Gx$.

Lemma: For any linear code $C \subseteq GF(q)^n$ we have

$$\boxed{d(C) = \min_{c \in C \setminus \{0\}} d(c)} .$$

proof: • let $c_1, c_2 \in C$ be such that $d(c_1 - c_2) = d(C)$.

$\tilde{c} := c_1 - c_2 \in C \setminus \{0\}$ then implies $d(C) = d(\tilde{c}) \geq \min_{c \in C \setminus \{0\}} d(c)$.

• conversely, if $c_1 \in C \setminus \{0\}$ s.t. $d(c_1) = \min_{c \in C \setminus \{0\}} d(c)$, then for $c_2 := 0$

$$d(C) \leq d(c_1 - c_2) = d(c_1) = \text{ -"- } .$$

□

Def.: A generator matrix $G \in GF(q)^{n \times k}$ is said to be "in systematic form" if $G = (\mathbb{1}_k \; P)^T$ for some $k \times (n-k)$ matrix P. The encoding $x \mapsto Gx$ is then also called "systematic".

remarks: • for every code with generator matrix G' we can by linear operations construct one with generator matrix G in sys. form s.t. the two codes are "equivalent" in the sense that their lengths, rates & min. distances coincides.

• the codewords of a syst. encoding contain the raw message in the first k components followed by $(n-k)$ symbols introducing redundancy.

Prop.: Let $G = (\mathbb{1}_k \; P)^T$ be the generator matrix of a linear code $C \subseteq GF(q)^n$. Then $\forall c \in GF(q)^n$:

$$\boxed{c \in C \Leftrightarrow Hc = 0} \quad \text{for } H := (-P^T \; \mathbb{1}_{n-k}) .$$

proof: • $c \in C \Rightarrow \exists x \in GF(q)^k : c = Gx$

$$\Rightarrow Hc = HGx = (-P^T \; \mathbb{1}) \begin{pmatrix} \mathbb{1} \\ P^T \end{pmatrix} x = (P^T - P^T) x = 0 \quad \checkmark$$

- $Hc = 0 \Rightarrow 0 = (-P^T \ \mathbb{1}) \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = c_2 - P^T c_1 \Rightarrow c_2 = P^T c_1$

$$\Rightarrow c = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} c_1 \\ P^T c_1 \end{pmatrix} = \begin{pmatrix} \mathbb{1} \\ P^T \end{pmatrix} c_1 = G c_1 \quad \checkmark$$

$\square$

remarks:
- if $c \in C$ is corrupted via $c \mapsto c' := c + e$, then $Hc' = He$ is independent of the original codeword.
- $H e$ is called "syndrome" & H is called "parity check matrix"
- a possible decoding strategy is then to infer/guess e from the syndrome.

V.3. Bounds on the performance of error correcting codes

Prop.: (Hamming bound) Let $C \subseteq \mathcal{X}^n$ be a code with $|\mathcal{X}| =: q$, distance $d(C) =: d$ and $m := \lfloor \frac{d-1}{2} \rfloor$ (= # of errors which can be corrected).

Then
$$\boxed{|C| \leq \frac{q^n}{V(q,n,m)}} \quad \text{where } V(q,n,m) := \sum_{i=0}^{m} \binom{n}{i} (q-1)^i$$

proof: For each $c \in C$ define a neighborhood $B_m(c) := \{ y \in \mathcal{X}^n \mid d(y-c) \leq m \}$.
Then $B_m(c) \cap B_m(c') = \emptyset$ for $c, c' \in C$ with $c \neq c'$ and $|B_m(c)| = V(q,n,m)$.

So $$|\mathcal{X}^n| = q^n \geq \left| \bigcup_{c \in C} B_m(c) \right| \underset{\uparrow}{=} \sum_{c \in C} |B_m(c)| = |C| \, V(q,n,m)$$

u(c)'s disjoint

$\square$

remark: if '=' holds in the Hamming bound, then we have a perfect packing of non-overlapping Hamming balls that cover the full space.

Def.: A code for which '=' holds in the Hamming bound is called „perfect".

Thm.: (Tietavainen / van Lint '70ies) The following are all perfect binary codes (i.e. $q=2$):

(i) $[2^r-1, 2^r-1-r, 3]$ Hamming codes (e.g. $[7,4]$ for $r=3$)

(ii) the "$[23,12,7]$ Golay code"

(iii) trivial codes (meaning $|C| \in \{1, 2^n\}$)

(iv) repetition codes $x_i \mapsto \underbrace{(x_i, \dots, x_i)}_{n \text{ times}}$ for odd n

Thm.: (Gilbert-Varshamov bound)

For every triple $(q,n,d) \in \mathbb{N}^3$ there exist a code $C \subseteq \mathcal{X}^n$ with $|\mathcal{X}| = q$ and distance $d(C) = d$ s.t.

$$\boxed{|C| \geq \frac{q^n}{V(q,n,d-1)}} \quad \text{where } V(q,n,d-1) = \sum_{j=0}^{d-1} \binom{n}{j} (q-1)^j$$

$$= \text{Volume of } B_{d-1} \in \mathcal{X}^n$$

proof: construct the code step-by-step via:

(i) start with arbitrary first codeword

(ii) add any point as a codeword which has Hamming distance at least d from all previously chosen codewords,

(iii) iterate (ii) until the Hamming balls of radius $(d-1)$ around the codewords cover all of $\mathcal{X}^n$.

The constructed code then satisfies $|C| \cdot V(q,n,d-1) \geq q^n$. □

remarks: • There are linear codes satisfying this bound. In fact, random linear codes do the job for large enough n.

- computing (even approximating) the distance of a linear code is NP-hard
  - → picking a random code & checking whether it has good distance is not feasible.
- for prime powers ≥ 49 there are explicit constructions based on algebraic geometry which satisfy the GV bound.
- for $q=2$ no explicit construction is known.

recall: $\circ$ $V(q,n,r) := \sum_{j=0}^{r} \binom{n}{j}(q-1)^j$ Volume of Hamming ball $B_r \subseteq \mathbb{Z}_q^n$

- Gilbert-Varshamov bound: $\forall (q,n,d) \in \mathbb{N}^3$ $\exists$ code $C \subseteq \mathbb{Z}_q^n$ s.t.

  $d(C) = d \ \wedge \ |C| \geq q^n / V(q,n,d-1)$

- $f(x) = o(g(x)) \iff \lim_{x\to\infty} \left|\frac{f(x)}{g(x)}\right| = 0$, e.g. $f(x) = o(1)$ means $f(x) \xrightarrow[x\to\infty]{} 0$.

- $f(x) = \Omega(g(x)) \iff \liminf_{x\to\infty} \left|\frac{f(x)}{g(x)}\right| > 0$ i.e. g is asympt. lower bound.

Lemma: For $p \in [0, \frac{1}{2}]$ and increasing $n \in \mathbb{N}$, we have

$$\boxed{2^{(h(p) - o(1))n} \leq V(2,n,pn) \leq 2^{h(p)n}}$$

(where $h(p) := -p\log p + (1-p)\log(1-p)$ is the binary entropy).

Corollary: (i) For $p \in [0, \frac{1}{2}]$ there is a sequence of binary codes $(C_n)_{n\in\mathbb{N}}$ with relative distance

$$\boxed{\begin{aligned} \frac{d(C_n)}{n} &\geq p \\ R(C_n) &\geq 1 - h(p) \end{aligned}}$$

such that.

(ii) Conversely, for $p \in [0,1]$ every sequence of binary codes with $\frac{d(C_n)}{n} \xrightarrow{n\to\infty} p$ satisfies

$$\boxed{R(C_n) \leq 1 - h\left(\tfrac{p}{2}\right) + o(1)}$$

proof: (i) by definition $R(C_n) := \frac{\log|C_n|}{n}$.

$R(C_n) \geq 1 - \frac{1}{n}\log V(2,n,d-1)$ by Gilbert-Varshamov

$\geq 1 - h(p)$ using the Lemma for $pn = d-1$

(ii) $R(C_n) \leq 1 - \frac{1}{n}\log V(2,n,\lfloor\frac{d-1}{2}\rfloor)$ Hamming bound

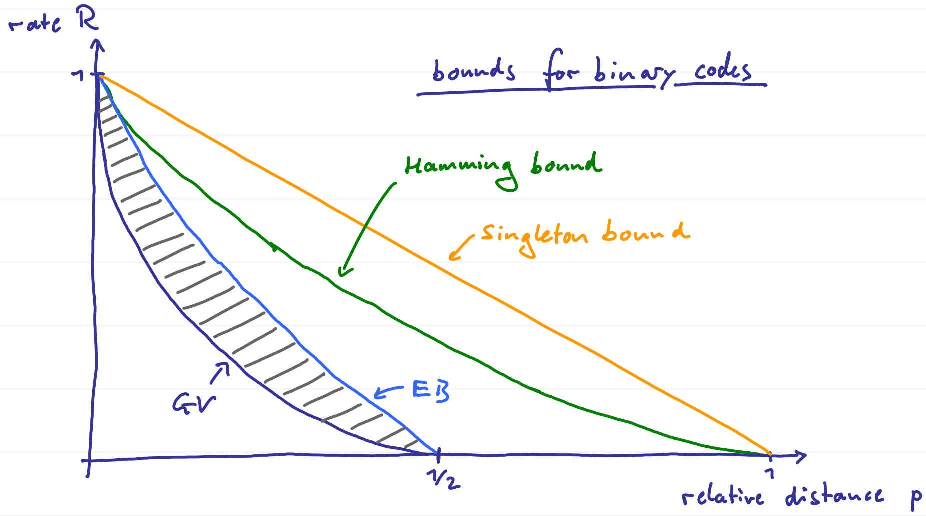
$\leq 1 - h\left(\frac{p}{2}\right) + o(1)$ Lemma

$\square$

Consequence: $\boxed{\text{Asymptotically good binary codes exist!}}$

remark: an improved upper bound is the "Elias-Bassalygo" bound:

$$\boxed{R(C_n) \leq 1 - h\left(\frac{1-\sqrt{1-2p}}{2}\right) + o(1)} \quad \text{for } n \to \infty$$



Prop.: (Singleton bound)

For every q-ary code $C \subseteq \mathbb{Z}_q^n$ with block length $n \in \mathbb{N}$ and distance d, we have $\boxed{|C| \leq q^{n-d+1}}$

proof:
- take all $|C|$ codewords and erase the first $(d-1)$ symbols
- we are left with $|C|$ strings which are distinct (since the distance was d) and of length $n-(d-1)$

$\Rightarrow \quad |C| \leq q^{n-d+1}$ = max # of q-ary strings of length $n-(d-1)$

$\square$

Corollary: Any linear $[n,k,d]$ code satisfies $\boxed{k \leq n-d+1}$

proof: $|C| = q^k$. $\square$

Def.: Linear $[n,k]$ codes with distance $d = n-k+1$ are called "maximum distance separable" (MDS) codes.

remarks: ∘ MDS codes require large alphabets:

a sequence of asymptotically good codes can be MDS only if $\boxed{q = \Omega\left(\frac{n}{\log n}\right)}$ with $n \to \infty$

∘ a sequence of MDS codes whose rel. distance is bounded away from 0 & 1 is asymptotically good, since

$$R = \frac{k}{n} \underset{\substack{\uparrow \\ \text{MDS}}}{=} 1 - \frac{d}{n} + \frac{1}{n}.$$

note that $R + \frac{d}{n} \to 1$ for MDS codes!

V. 4. Reed-Solomon codes

Def.: For integers $1 \leq k < n \leq q$ and $\alpha \in GF(q)^n$ with distinct components

$$C := \{ p(\alpha) \in GF(q)^n \mid p \text{ is polynomial over } GF(q) \text{ of degree } < k \}$$

is called "Reed-Solomon code" and we will write $[n,k]$-RS code.

Encoding: ∘ we identify a message $m \in GF(q)^k$ with a polynomial

$$p_m(x) := \sum_{l=0}^{k-1} m_l x^l$$

the codeword is then $p_m(\alpha) = (p_m(\alpha_1), \dots, p_m(\alpha_n))$

• $p_m(\alpha) = G m$ where $G \in GF(q)^{n \times k}$ is a "Vandermonde matrix" with $G_{xy} := \alpha_x^{y-1}$

$\Rightarrow$ RS codes are linear

remarks: ∘ RS codes have large alphabet (since $q \geq n$)

∘ typical choices for α:

(i) $\{\alpha\} = \{GF(q)\}$ i.e. $q = n$

(ii) $\{\alpha\} = \{GF(q)\} \setminus \{0\}$ i.e. $q = n+1$

$\alpha = (\beta^0, \beta^1, \dots, \beta^{n-1})$, β "primitive element" of $GF(q)$

Def.: $\mathbb{K}_d[X]$:= space of polynomials over the field $\mathbb{K}$ with degree $\leq d$.

Lemma: If $\alpha \in \mathbb{K}^{d+1}$ has distinct components, then $\hat{\alpha} : \mathbb{K}_d[X] \to \mathbb{K}^{d+1}$,

$\hat{\alpha} : p \mapsto (p(\alpha_1), \dots, p(\alpha_{d+1}))$ is bijective.

proof: "Lagrange interpolation" define $L_i \in \mathbb{K}_d[X]$,

$$L_i(x) := \frac{\prod_{k \neq i} (x - \alpha_k)}{\prod_{l \neq i} (\alpha_i - \alpha_l)} \quad , \quad i, k, l \in \{1, \dots, d+1\}$$

Then $L_i(\alpha_j) = \delta_{ij}$. For any $\beta \in \mathbb{K}^{d+1}$ define $p(x) := \sum_{i=1}^{d+1} \beta_i L_i(x)$.

Then $p \in \mathbb{K}_d$ and $p(\alpha_i) = \beta_i$. Hence $\hat{\alpha}$ is surjective.

Conversely, if $p, \tilde{p} \in \mathbb{K}_d[X]$, then $(p - \tilde{p}) \in \mathbb{K}_d[X]$ has $(d+1)$ roots $\{\alpha_i\}_{i=1}^{d+1}$

$\Rightarrow$ $p - \tilde{p} = 0$, so $\hat{\alpha}$ is also injective.

$\square$

Thm.: For an $[n,k]$-RS code over $GF(q)$ we have

(i) $|C| = q^k$

(ii) $d(C) = n-k+1$

proof: (i) follows from injectivity of G

(ii) Linearity $\Rightarrow$ $d(C) = \min_{c \in C \setminus \{0\}} d(c)$

for $m \in GF(q)^k \setminus \{0\}$ $p_m(x)$ has at most $k-1$ roots as $p_m \in \mathbb{K}_{k-1}[X]$.

$\Rightarrow$ codeword $c = p_m(\alpha)$ has at most $k-1$ zeros

$\Rightarrow$ $d(C) \geq n-k+1$

Singleton bound: $d(C) \leq n-k+1$.

$\square$

Corollary: RS-codes are MDS codes (i.e., they achieve the Singleton bound)

V.5. Error bursts & interleaving

sources for errors are often not memoryless / uncorrelated, e.g.:

- scratches on CD
- disturbance / loss of signal for time intervals

→ "bursts" of errors

simple ways to deal with this:

(i) use codes with large alphabet (e.g. RS) & represent symbols using smaller alphabets. E.g. code over $GF(2^m)$ with distance d corrects bursts of length $\left(\lfloor \frac{d-1}{2} \rfloor - 1\right) m + 1$ when information is stored using contiguous bits.

(ii) interleaving = rearranging symbols in concatenated codewords.

Consider $[n,k]$-code & let $c^{(i)} = (c_1^{(i)}, \dots, c_n^{(i)}) \in C$, $i \in \{1, \dots, t\}$.

Define new $[nt, kt]$-code $\tilde{C}$ from all codewords of the form

$$\tilde{c} = \left(c_1^{(1)} c_1^{(2)} \cdots c_1^{(t)} c_2^{(1)} \cdots c_2^{(t)} \cdots c_n^{(t)} \right)$$

C corrects bursts of length $b \Rightarrow \tilde{C}$ corrects bursts of length $\tilde{b} = t \cdot b$

Example: $[256, 229]$ - RS code: rate $\sim 90\%$, corrects 13 byte errors

→ corrects bursts of $12 \cdot 8 + 1 = 97$ bit errors

→ $t = 31$ interleaving corrects burst up to 3 kbits

(essentially this happens on a CD. 3 kbits $\hat{=}$ 2.5 mm on surface)

SIGNAL RECOVERY & UNCERTAINTY RELATIONS

notation:

$\mathcal{B}(\mathbb{R}) :=$ Borel sets on $\mathbb{R}$

For $T \in \mathcal{B}(\mathbb{R})$ $|T| := \int_T dt$ Lebesgue measure of T

$f: \mathbb{R} \to \mathbb{C}$ signal in the time domain

$\|f\|_p := \left(\int |f(t)|^p dt\right)^{1/p}$, $L^p :=$ equivalence class of functions with $\|f\|_p < \infty$

$\hat{f}(\omega) := \int_{-\infty}^{\infty} f(t)\, e^{-2\pi i \omega t}\, dt$ Fourier transformed signal (frequency domain)
(recall Parseval: $\|\hat{f}\|_2 = \|f\|_2$)

For $A: L^p \to L^q$: $\|A\|_{p\to q} := \sup_{f \in L^p \setminus \{0\}} \frac{\|Af\|_q}{\|f\|_p} = \|A\|$ if clear from context

Def.: • For $T, W \in \mathcal{B}(\mathbb{R})$ let $P_W, P_T: L^2(\mathbb{R}) \to L^2(\mathbb{R})$ be the "time-limiting" & "frequency-limiting" operators defined as $P_T f(t) := \begin{cases} f(t), & t \in T \\ 0 \end{cases}$ and

$P_W f(t) := \int_W e^{2\pi i \omega t}\, \hat{f}(\omega)\, d\omega$ (densly defined on $L^2(\mathbb{R})$)

• f is said to be "ε-L^p-concentrated" on $T \in \mathcal{B}(\mathbb{R})$ iff $\|f - P_T f\|_p \le \varepsilon \|f\|_p$

• $\hat{f}$ is "ε-L^p-concentrated /band-limited" on $W \in \mathcal{B}(\mathbb{R})$:iff $\|f - P_W f\|_p \le \varepsilon \|f\|_p$

(note that by Parseval's identity: $\|f - P_W f\|_2 = \|\hat{f} - \widehat{P_W f}\|_2$)

Lemma: $\|P_W P_T\|^2_{2\to 2} \le |W| \cdot |T|$

proof (sketch): note
$$P_W P_T f(s) = \int_W e^{2\pi i \omega s} \int_T e^{-2\pi i \omega t} f(t)\, dt\, d\omega$$
$$= \int_T \int_W e^{2\pi i (s-t)\omega}\, d\omega\, f(t)\, dt$$
$$=: \int_{\mathbb{R}} k(s,t)\, f(t)\, dt \Rightarrow P_W P_T \text{ is compact operator}$$
$$\Rightarrow \|P_W P_T\|_{2\to 2} \le \underset{\substack{\uparrow \\ \text{Schatten 2-norm}}}{\|P_W P_T\|_2} = \int_T \int_W d\omega\, dt = |T| \cdot |W|$$
$\square$

Theorem: [L^2- uncertainty relation]

Let $T, W \in \mathcal{B}(\mathbb{R})$ and f and $\hat{f}$ be ε_T and ε_W L^2-concentrated on T and W respectively. Then

$$\sqrt{|W|\cdot|T|} \geq \| P_W P_T \|_{2\to 2} \geq 1 - (\varepsilon_T + \varepsilon_W)$$

proof:

- $$\| f - P_W P_T f \| \underset{\Delta \text{ ineq.}}{\leq} \| f - P_W f \| + \underbrace{\| P_W (f - P_T f) \|}_{\leq \underbrace{\|P_W\|}_{=1} \|f - P_T f\|} \leq \varepsilon_W + \varepsilon_T$$

- $$\| f - P_W P_T f \| \geq \| f \| - \| P_W P_T f \|$$

$$\Rightarrow \frac{\| P_W P_T f \|}{\| f \|} \geq 1 - \varepsilon_T - \varepsilon_W$$

$$\frac{\| P_W P_T f \|}{\| f \|} \leq \| P_W P_T \| \underset{\text{Lemma}}{\leq} \sqrt{|W||T|}$$

$\square$

Recovery of missing segments:

- Assume $f \in L^2(\mathbb{R})$ is W-band-limited in the sense that $P_W f = f$
- Let $\eta \in L^2(\mathbb{R})$ be additive noise to f, i.e. $f \mapsto f + \eta$
- Assume the signal is missing in a time window T

$\rightarrow$ finally received signal is $\phi := (\mathbb{1} - P_T)(f + \eta)$

Thm.: If $\| P_W P_T \| < 1$ (i.e. in particular if $|W|\cdot|T| < 1$), then there is a recovery operator $R : L^2 \to L^2$ s.t.

$$\boxed{\| f - Rf \|_2 \leq \frac{\| \eta \|_2}{1 - \| P_W P_T \|}}$$

proof.: Define $R := (\mathbb{1} - P_T P_W)^{-1}$ and note that $\|P_W P_T\| = \|P_T P_W\|$

Then $\|f - R\phi\|_2 = \|f - R(\mathbb{1} - P_T)(f + \eta)\|_2$

$$\overset{f = P_W f}{=} \|f - f - R(\mathbb{1} - P_T)\eta\|_2$$

$$= \|R(\mathbb{1} - P_T)\eta\|_2 \leq \|R\| \underbrace{\|\mathbb{1} - P_T\|}_{=1} \|\eta\|_2$$

Moreover $\|R\| = \|(\mathbb{1} - P_T P_W)^{-1}\| \leq (1 - \|P_T P_W\|)^{-1}$, so that

$$\|f - R\phi\|_2 \leq \frac{\|\eta\|}{1 - \|P_T P_W\|}$$ $\square$

remark: $R = (\mathbb{1} - P_T P_W)^{-1} = \sum_{k=0}^{\infty} (P_T P_W)^k$ suggests a recovery algorithm making use of alternating projections.

Thm.: [L^1-uncertainty relation] If f is ε_T-L^1 concentrated on T & band limited on W, then $\boxed{|W| \cdot |T| \geq 1 - \varepsilon_T}$

proof: • by hopothesis $\frac{\|P_T f\|_1}{\|f\|_1} \geq 1 - \varepsilon_T$

• for f L^1 bandlim. it holds that $\|f\|_\infty \leq |W| \, \|f\|_1$

• on the other hand: $\|P_T f\|_1 = \int_T |f(t)| \, dt \leq \|f\|_\infty |T|$

$$\Rightarrow \frac{\|P_T f\|_1}{\|f\|_1} \leq |T| \cdot |W|$$ $\square$

Correction of sparse noise:

Assume a band-limited signal $f = P_W f$ is sent over a "noisy channel" which adds "sparse noise" (= supported on T) so that the received signal is

$$\phi = f + P_T \eta . \qquad \text{(no bound on } \|\eta\| \text{!)}$$

With $B_1(W) := \{ f \in L^1(\mathbb{R}) \cap L^2(\mathbb{R}) \mid \|f\|_1 = 1 \wedge P_W f = f \}$ we get

Thm.: (Logan's phenomenon)

$$|W| \cdot |T| < \tfrac{1}{2} \Rightarrow f = \underset{\varphi \in B_1(W)}{\operatorname{argmin}} \|\varphi - \phi\|_1$$

proof: Since $|W||T_\varphi| \geq 1$ for any $\varphi \in B_1(W)$ with support T_φ

$|W||T| < \frac{1}{2}$ means that $\|P_T \varphi\|_1 < \frac{1}{2}\|\varphi\|_1$ and so

$$\|P_T \varphi\|_1 < \|P_{T^c} \varphi\|_1 \quad (\text{since } \|\varphi\|_1 = 1)$$

Therefore the best bandlimited approximation to η is zero since:

for $\varphi = P_W \varphi$:

$$\|\eta - \varphi\|_1 = \|P_T(\eta - \varphi)\|_1 + \|P_{T^c}(\eta - \varphi)\|_1$$

$$\underset{\Delta \text{ ineq. } \& \ P_{T^c}\eta = 0}{\geq} \|P_T \eta\|_1 - \|P_T \varphi\|_1 + \|P_{T^c} \varphi\|_1$$

$$> \|P_T \eta\|_1 = \|\eta\|_1$$

To prove the Thm. suppose $f \neq 0$ & note that $\|f + \eta - \varphi\|_1 = \|\eta - \underbrace{(\varphi - f)}_{\text{band limited}}\|_1$ is minimized for $\varphi = f$. $\square$

note:
- T is unknown here (we just make use of small $|T|$)
- generalizations in various directions in the compressed sensing community