07/07/2017 GUILHEM: BRIDGES BETWEEN INFORMATION THEORY AND STAT MECH -

## 1 INTRODUCTION:

Nowadays, manipulating daily loads of info: images, text, sounds = files = 011000
On which we are performing : - compression (file) $\rightarrow$ (smaller file) $\simeq$

- transmission over noisy channel (file) $\xrightarrow{\text{noise}}$ (computer file) error correction

These two problems were already formalised and approximately solved by Shannon 48 "A mathematical theory of communication".

Why should we care about this while not being telecommunication engineer?
bio applicat°: - DNA $\rightarrow$ RNA $\rightarrow$ proteins
- vision: receptors/ retina $\rightarrow$ optimal nerve $\rightarrow$ brain
social: - language eg: - repetition $\rightarrow$ error correcting.
- only some combination of sounds make up words

theoretically: Information theory enlightens statistical mechanics.
$\rightarrow$ fruitful interdisciplinary of stat mech of disordered systems. (Cris, Florent, Federico).
e.g. Constraint satisfaction problems for error corrections LDPC

↳ WEB PAGE: www.phys.ens.fr/~guilhem/boulder.html.

## 2 THE MEANING OF ENTROPY:

2A definition: $S_{\mu c} = k \ln \Omega$
microcanonical

$$S_c = -k \sum_{\mathcal{C}} p(\mathcal{C}) \ln p(\mathcal{C}) \quad \text{with } p(\mathcal{C}) = \frac{1}{Z} e^{-\beta H \mathcal{C}}$$

More IT:
↳ Shanon entropy for distribution $p = \{ p(x) ; x \in \underset{\text{alphabet}}{\mathcal{X}} \} \equiv H(p) = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x)$
(change of unit of Boltzmann $k_B$).

prop: $H(p) = 0 \iff p(x) = \delta_{x, x_0}$ concentrated on one value

$H(p) \in [0, \log_2 |\mathcal{X}|]$ and $H(p) = \log_2(|\mathcal{X}|)$ maximal $\iff p(x) = \frac{1}{|\mathcal{X}|} \; \forall x$ -

"the entropy measures the lack of information on a realization of $p$"

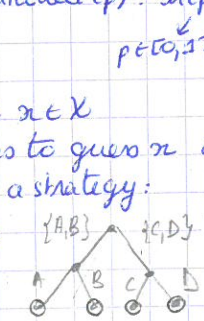
↳ yet there are many functions growing between uniform and concentrated, so why should it be $H(p)$ with the logarithm?

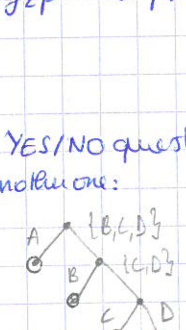

definition for bernoulli$(p)$: $h(p) = -p \log_2 p - (1-p) \log_2 (1-p)$
$p \in [0,1]$

2B guessing game: { Amy choose $x \in \mathcal{X}$
Sheldon has to guess $x$ asking YES/NO questions as fast as possible.

a strategy: another one:

e.g. $\mathcal{X} = \{A, B, C, D\} \rightarrow$



formalisation: $T$ = strategy = tree of questions
$\ell_x(T)$ = # questions to find $x \in \mathcal{X}$ with strategy $T$

↳ to define the best strategy we compute $\bar{\ell}(T) = \sum_{x \in \mathcal{X}} p(x)\, \ell_x(T) \Rightarrow T^* = \underset{T}{\operatorname{argmin}}\, \bar{\ell}(T)$

"$T^*$ minimizes the number of questions Sheldon expects to ask"

this is bounded: $\boxed{H(p) \leqslant \bar{\ell}(T^*) \underset{\text{true with } \leqslant}{\leqslant} H(p)+1}$ SHANON SOURCE CODING THM

↳ measures the lack of information within one bit of $x$ knowing $p$ !!

a proof: - Kraft inequality $\forall T, \sum_{x \in \mathcal{X}} 2^{-\ell_x(T)} \leqslant 1$

↳ $\ell_{max} = \max_{x \in \mathcal{X}} \ell_x(T)$


total number leaves (at $\ell max$) $2^{\ell max} \geqslant \sum_{x \in \mathcal{X}} 2^{\ell max - \ell(x)}$ (total number of shadows) → Kraft. ∎

- Kullback-Leibler divergence: $p, q$ proba laws on $\mathcal{X}$ $\left\{\begin{array}{l} p(x) \in [0,1] \\ \sum_x p(x) = 1 \end{array}\right.$ $D(p\|q) = \sum_{x \in \mathcal{X}} p(x) \log_2 \frac{p(x)}{q(x)}$

prop: - $D(p\|q) \geqslant 0$ (due to the concavity of the log)
- $D(p\|q) = 0 \iff p = q$
} CAN BE THOUGHT AS A DISTANCE.

- the proof:
* consider $q(x) = \frac{1}{z}$ with $z = \sum_{x \in \mathcal{X}} 2^{-\ell_x} \leqslant 1$ by Kraft

↳ $D(p\|q) = \sum_x p(x) \log_2 \frac{p(x)}{2^{-\ell_x}} z = -H(p) + \log_2 z + \bar{\ell}(T) \geqslant 0$

$\Rightarrow H(p) \leqslant \bar{\ell}(T)$

* if $\{\ell_x\}_{x \in \mathcal{X}}$ integers s.t. $\sum_x 2^{-\ell_x} \leqslant 1 \Rightarrow \exists T$ such that $\ell_x(T) = \ell_x \ \forall x$.

↳ $\ell_{x_1} \leqslant \ell_{x_2} \leqslant \dots$ put $\left\{\begin{array}{l} x_1 \text{ in the leftmost node of depth } \ell_{x_1} \\ x_2 \text{ in the remaining leftmost node of depth } \ell_{x_2} \end{array}\right.$

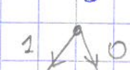
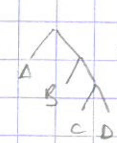
we would like to take $\ell_x = \left\lceil \log_2 \frac{1}{p(x)} \right\rceil$ → obey Kraft inequality

$\leqslant \log_2 \frac{1}{p(x)} + 1$

$\bar{\ell}(T) \leqslant \sum_x p(x) \left[-\log_2 p(x) + 1\right] \leqslant H(p) + 1$ ← $\exists T$ !

**2C data compression:**

consider a file of string of symboles in $\mathcal{X}$: $x_1 x_2 x_3 \dots \longrightarrow 00101\dots$ as short as possible

It is actually the problem we have solve:  so that  ≡ A = 0
B = 10
C = 110
D = 111

let's formalise that with $w: x_i \longrightarrow w_{x_i}$
stacked strings of words built in these way are uniquely decodable, (prefix free
going back to the root until reading word once and once — on the fl..)

if $x_i$'s are iid with proba $p(x)$ $\rightarrow$ need $nH(p)$ bits to compress $x_1 \dots x_n$
$\equiv$ "entropy is the best rate of compression possible".

## 2D mutual information

An inference problem: q ess of random $X$ from observation $Y$.
formally: the pair of r.v $(X, Y) \in (\mathcal{X}, \mathcal{X}')$
distributed according to $p_{X,Y}(x,y) = \mathbb{P}[X = x, Y = y]$

$\rightarrow$ marginal laws: $p_X(x) = \sum_y p_{X,Y}(x,y)$, $p_Y(y) = \sum_x \dots$

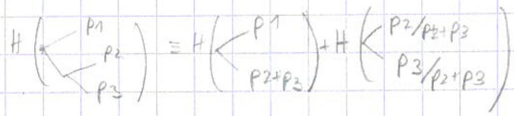
$\rightarrow$ conditional laws: $p_{X|Y}(x|y) = \dfrac{p_{X,Y}(x,y)}{p_Y(y)}$

From which we consider various interesting entropies:

* joint law entropy $H(X,Y) = -\sum_{x,y} p_{X,Y}(x,y) \log_2 p_{X,Y}(x,y)$
* entropy of marginal $H(X) = -\sum_x p_X(x) \log_2(p(x))$
* conditional ent. $H(X|Y) = -\sum_y p_Y(y) \sum_x p_{X|Y}(x|y) \log_2 p_{X|Y}(x|y)$
$= -\sum_{x,y} p_{X,Y}(x,y) \log_2 p_{X|Y}(x|y)$
* mutual info. $I(X,Y) = D(p_{X,Y} \,\|\, p_X p_Y)$ $\Rightarrow$ prop: $I(X,Y) = 0 \Leftrightarrow X, Y$ indep.
$= \sum_{x,y} p_{X,Y}(x,y) \log \dfrac{p_{X,Y}(x,y)}{p_X(x) p_Y(y)}$
$= H(X) + H(Y) - H(X,Y)$ $\Rightarrow$ prop: $H(X,Y) \leq H(X) + H(Y)$
$= \sum_{x,y} p_{X,Y}(x,y) \log \dfrac{p_{X|Y}(x|y)}{p_X(y)}$
$= H(X) - H(X|Y)$ $\Rightarrow$ prop: $H(X|Y) \leq H(X)$
"conditionning reduce entropy"

$\rightarrow$ EXERCIZES: - show that Gibbs maximize entropy under constraint $\langle E \rangle = U$
- proove that uniform distribution maximises entropy under no-constraint (KL div)
- SHANNON'S PAPER $\rightarrow$ Appendix 2:

$H(p_1, \dots, p_M)$ continuous in $p_i$
$H(\frac{1}{n}, \dots \frac{1}{n})$ growing with $n$ $\Rightarrow H = -\text{cte} \sum_{i=1}^{n} p_i \log p_i$

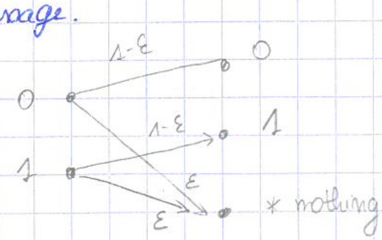



# 3 COMMUNICATION OVER NOISY CHANNELS

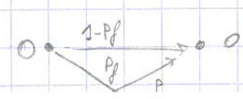
## 3A definitions

message $\xrightarrow{\text{noise}}$ corrupted message.

examples: $\rightarrow$ Binary Erasure Channel: (BEC)
$\varepsilon \equiv$ probability of erasure.




$\rightarrow$ Binary Symmetric Channel: (BSC)

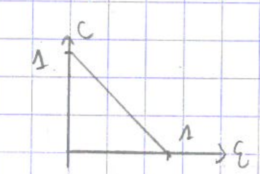



in which case you can never be entirely sure

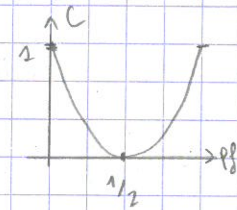
10/07/2017

" the higher, the better you can deduce input from output "

**Capacity of a channel:** $C = \max_{P_X} I(X;Y)$ with $X$ input of a channel.
$Y$ output.
$P_X$ probability law of input.




EXERCIZES: $\begin{cases} C_{BEC} = 1-\varepsilon \\ C_{BSC} = 1-h(p_f) \end{cases}$




**Encoding and decoding:** message $\xrightarrow{\text{ENCODER}}$ encoded message $\xrightarrow[\text{CHANNEL}]{\text{NOISY}}$ corrupted encoded message $\xrightarrow{\text{DECODER}}$ decoded message ?

"string of bit" "longer string"

**Rate of a code:** $\equiv \dfrac{\#\text{ bits of message}}{\#\text{ bits of encoded msg.}} < 1$ → the larger the better to reduce the cost of message transmission.

## 3B naive coding ≡ repetition

e.g. we repeat 3 times input.

ENCODER
$0 \longrightarrow 0\,0\,0$ with the BSC $p_f < 1/2$.
$1 \longrightarrow 1\,1\,1$

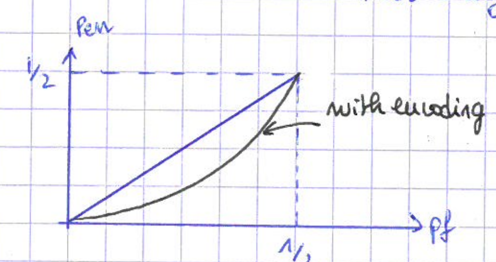
reasonable DECODER
"majority rule"
↳ odd number of bits

$000 \longrightarrow 0$
$111 \longrightarrow 1$
$001, 010, 100 \longrightarrow 0$
$110, 101, 011 \longrightarrow 1$

rk: rate $= 1/3$.

How good is the naive coding?

Probability of error without encoding $= p_f$
Probability of error with 3 repetitions $= p_f^3 + 3p_f^2(1-p_f)$
3 flips; $\binom{3}{2}$ flips.




→ better than no encoding but
$\begin{cases} P_{en} > 0 \text{ as soon as } p_f > 0 \\ \text{rate} = 1/3 \end{cases}$
↳ actually not that good. Can one do better?

## 3C Shannon channel coding theorem

$\boxed{\text{THM: There exist encoding (of growing length) with } P_{en} \xrightarrow{N\to\infty} 0, \text{ for all rates } R \text{ smaller than the capacity of the channel } C \ (R<C).}$

So that we can now interpret the capacity as the best acheivable rate with $P_{en} \to 0$.
The paradoxical statement that we could be sure of the signal despite the noise its Resolved by the fact this is an asymptotic statement (thermodynamic limit).

**MORE FORMAL DEFINITIONS:**

message: $\underline{z} = (z_1, \dots z_L) \in \{0,1\}^L$

encoded message: $\underline{x} = (x_1, \dots x_N) \in \{0,1\}^N$

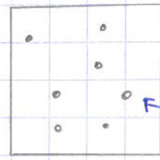
corrupted encod. msg.: $\underline{y} \in \mathcal{X}_{out}^N$
BEC $\mathcal{X}_{out} = \{0,1,N\}$
BSC $\mathcal{X}_{out} = \{0,1\}$

corrected encoded msg: $\hat{\underline{x}}(\underline{y})$

corrected msg: $\hat{\underline{z}}(\hat{\underline{x}}(\underline{y}))$

encoding: $\underline{x} = f_{\text{encoding}}(\underline{z})$

Code book $= \mathcal{C} = \{\underline{x}^{(1)}, \dots, \underline{x}^{(2^L)}\}$



$\{0,1\}^N$ exponentially many but exponentially few among available code words

$2^L$ codewords $\{\underline{x}^{(1)}, \dots \underline{x}^{(2^L)}\}$

channel:

$\underline{y}$ $\underline{x}^{(\alpha)}$ in $\{0,1\}^N$

intuition → put the codewords as far as possible in space BUT! CAREFUL IN HIGH DIMENSIONS, things do not happen the same way as in 2d!! oversimplifying view.

correction: $\hat{\underline{x}}(\underline{y})$ estimation of $\underline{x}^{(\alpha)}$

decoding: $\hat{\underline{z}} = f^{-1}_{\text{encoding}}(\hat{\underline{x}}(\underline{y}))$

→ From now on we will focus on the central $\underline{x} \to \underline{y} \to \hat{\underline{x}}(\underline{y})$. If the $2^L$ $\underline{z}$ are uniform at random, the $\underline{x}$ are also u.a.r.

DECODING AS AN INFERENCE PROBLEM:

$\underline{X}$ random code word, $P_{\underline{X}}(\underline{x}) = 1/2^L \; \mathbb{1}(\underline{x} \in \mathcal{C})$

$\underline{Y}$ random output $P_{\underline{Y}|\underline{X}}(\;|\;) = \prod_{i=1}^{N} Q(y_i|x_i)$

over the bits indep.

BEC: $\begin{cases} Q(0|0) = 1-\varepsilon \\ Q(*|0) = \varepsilon \\ Q(1|0) = 0 \end{cases}$

using Bayes theorem: $P_{\underline{X}|\underline{Y}}(\underline{x}|\underline{y}) = P_{\underline{Y}|\underline{X}}(\underline{y}|\underline{x}) \dfrac{P_{\underline{X}}(\underline{x})}{P_{\underline{Y}}(\underline{y})}$ ← prior proba on the signal

$= \prod_i Q(y_i|x_i) \dfrac{1}{2^L} \mathbb{1}(\underline{x} \in \mathcal{C}) \dfrac{1}{P_{\underline{Y}}(\underline{y})}$

$= \dfrac{1}{Z(\underline{y})} \mathbb{1}(x \in \mathcal{C}) \prod_{i=1}^{N} Q(y_i|x_i)$ POSTERIOR PROBABILITY

one keeps only the factors function of $\underline{x}$ and lets the rest into the partition $Z(\underline{y})$.

Rk: Having all the distribution of the decoding given the observation implies that we have ≠ ways of decoding:

* $\hat{\underline{x}}(\underline{y}) = \underset{\underline{x}}{\operatorname{argmax}} \; P_{\underline{X}|\underline{Y}}(\underline{x}|\underline{y})$ block MAP (Maximal a posteriori) → $\min P(\hat{\underline{X}} \neq \underline{X})$
* $\hat{x}_i(\underline{y}) = \underset{x_i}{\operatorname{argmax}} \; P_{X_i|\underline{Y}}(x_i|\underline{y})$ symbol MAP decoding → $\min \mathbb{E}(d(\hat{\underline{X}}, \underline{X}))$

EXAMPLE WITH THE BEC: $\underline{y} = (0, 1, 0, 0, *, *, 0 \dots *, 0)$ (erase)

$P_{\underline{X}|\underline{Y}}(\underline{x}|\underline{y}) = \dfrac{1}{|\mathcal{C} \cap B(\underline{y})|} \mathbb{1}(\mathcal{C} \cap B(\underline{y}))$ intersection Codbook and ball

$B(\underline{y}) = \{\underline{x} \in \{0,1\}^N \text{ s.t. } y_i \in \{0,1\} \Rightarrow x_i = y_i\}$

uniform probability over the messages matching the revealed bits in received message.

How should we construct code books? Let's start with a simple construction proposed by Shannon.

SHANON RANDOM CODE ENSEMBLE ↗ connection to the REM!

$\boxed{\mathcal{C} = \{\underline{x}^{(1)}, \dots \underline{x}^{(2^L)}\} \text{ with } \underline{x}^{(\alpha)} = \{x_1^{(\alpha)}, \dots x_N^{(\alpha)}\} \to \text{choose the } N \times 2^L \; x_i^{(\alpha)} \text{ iid } 0, 1 \text{ with probability } 1/2.}$

Rk: proba that $\underline{x}^{(\alpha)} = \underline{x}^{(\beta)}$ for $\alpha \neq \beta$ $\xrightarrow[L \to +\infty]{N \to +\infty} 0$ $R = \frac{L}{N}$ fixed. (non-injective)

Bin$(N, 1-\varepsilon)$ fluctuations

ANALYSIS ON THE BEC:

Assume w.l.o.g that $\underline{x}^{(1)}$ has been sent. $\underline{y} = \begin{cases} \underline{x}^{(1)} \text{ on } N(1-\varepsilon) + \cancel{O(\sqrt{N})} \text{ correctly transmitted} \\ * \text{ on other bits } N\varepsilon \end{cases}$

$W \overset{d}{=} \mathrm{Bin}\left(2^L - 1, \left(\frac{1}{2}\right)^{N(1-\varepsilon)}\right)$ "number of confusing code word".

(d: in distribution; $2^L - 1$: total # of available code words; $\left(\frac{1}{2}\right)$: proba 2 ≠ codewords agree on a bit)

$\Rightarrow E(W) = 2^{N(R-1+\varepsilon)} \longrightarrow 0$ if $\varepsilon < 1 - R$ $\Rightarrow P_r[W \neq 0] \longrightarrow 0$.

($W$: integer! First moment / Markov inequality for strictly positive code word.)

So that no confusion if $\underline{R < 1 - \varepsilon = C_{BEC}}$ → Shannon theorem! ∎

Even the random codes achieve capacity. (here we considered averages for codewords etc...)

HOMEWORK: proof for the BSC (text on website).

Rk: we proove that there exist such codes, but we did not show that we could not do better than this bound. if $R > C \Rightarrow P_{err} > 0$ coming from $H(\underline{X}|\underline{Y}) > 0$, Fano inequality.

⚠ In practice, encoding and decoding the Random Code ensemble takes exponential time and memory. $\mathcal{C}$ has to be descibed by $N 2^L$ bits. The encoding is easy by looking up the table. But the decoding is NP hard as one need to look for the corresponding match in the table. The randomness hinders any compression, we can only do exhaustive search in the table.

## 4 LOW DENSITY PARITY CHECKS CODES (LDPC). → putting some structure :)

### 4A Linear codes

- $\{0,1\}^N$ is a linear space over $\mathbb{Z}_2 = \{0,1\}$ "scalars", addition mod 2, multiplication
- The codebook $\mathcal{C} \subset \{0,1\}^N$ is said to be a linear code if $\mathcal{C}$ is a linear subspace of $\{0,1\}^N$.

$\underline{x} + \underline{y} = (\underbrace{x_1 + y_1}_{\text{mod } 2}, x_2 + y_2, \dots x_N + y_N) \in \mathcal{C}$

$\underline{x} + \underline{x} = \underline{0} \in \mathcal{C}$ (the origin belongs to the linear subspace).

↳ $\underline{0}$ is always a codeword of a linear code

All codwords are equivalent (can always make gauge transformation to change position of origin).