Bridges between statistical mechanics and information theory

Guilhem Semerjian

I Introduction

why these lectures, why in this school

- information everywhere, "big data" era

  ↘ texts, sounds, images ... = computer files = 0110 ...

  2 basic problems: . compressing these files ← how much information in it? (no moral prejudice)
  . transmitting through noisy channels (interference of cell phones)

  formalized by Shannon, 48, A mathematical theory of communication, fundamental paper

- other forms of information treatment

  - DNA $\xrightarrow[\text{transcription}]{}$ RNA $\xrightarrow{\text{translation}}$ proteins → cells → organs → organisms , lot of noise, regulations
  - photons → retina → optical nerve → brain perception
  - language, possibility of error corrections

- relationship with the school

  - enlightens fundamental features of stat mech (entropy)
  - one of the fruitful interdisciplinary applications of stat mech of disordered systems (noise = disorder)
    LDPC, CSPs

  → relations with C. Moore
  G. Biroli
  H. Cohn (packing in Hamming space)
  F. Krzakala
  F. Ricci-Tersenghi

deviations from original title of the lecture, but we'll see at the end a form of "cavity computation", to be completed by Federico

bibliography,
outline of the lectures,
exercises on:

www.phys.ens.fr/~guilhem/boulder.html

II The meaning of entropy

II. A. Definition

in stat. mech. $S = k \ln \Omega$ microcanonical

$$S = -k \sum_{\mathcal{C}} p(\mathcal{C}) \ln p(\mathcal{C}) \quad \text{canonical}, \quad p(\mathcal{C}) = \frac{e^{-\beta H(\mathcal{C})}}{Z}$$

if $p(\mathcal{C}) = \begin{cases} \frac{1}{Z} & \text{on } \Omega \text{ configurations} \\ 0 & \text{o.w.} \end{cases}$ reduces to microcanonical entropy

Shannon's definition and notation:

$p = \{ p(x), x \in \mathcal{X} \}$ probability law

$H(p) = - \sum_{x \in \mathcal{X}} p(x) \log_2 p(x)$ choice of units of $k = \frac{1}{\ln 2}$

$0 \ln 0 = 0$

"entropy is a measure of (the lack of) information"

$p(x) = \delta_{x, x_0} \iff H(p) = 0$

$p(x) = \frac{1}{|\mathcal{X}|} \iff H(p) = \log_2 |\mathcal{X}|$ maximal (exercise, to be done with $D(p \| q)$)

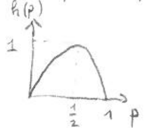
and in general $H(p) \in [0, \log_2 |\mathcal{X}|]$

$\Rightarrow$ grows when p spreads out, ie when randomness $\nearrow$, but

why this precise form? many other functions could be used

we shall see that this is the "right" def.

if $|\mathcal{X}| = 2$, $h(p) = -p \log_2 p - (1-p) \log_2 (1-p)$



II. B. A guessing game

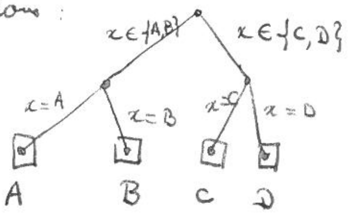
to be precised later

2 players, Alice (Amy) draw r.v. $X \in \mathcal{X}$, proba $p(x) = \mathbb{P}(X = x)$ → outcome x

Bob (Sheldon) does not see x, ask yes/no questions to Alice to determine x

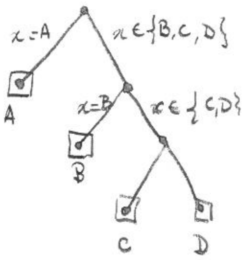
example: $\mathcal{X} = \{A, B, C, D\}$

tree of questions:



$\ell_A = \ell_B = \ell_C = \ell_D$, ℓ_x: nb of questions to determine x

but Bob can also play with:



$\ell_A = 1$, $\ell_B = 2$, $\ell_C = \ell_D = 3$

what's the best strategy? ie for Bob to conclude as fast as possible

it depends on $p(x)$, if A is very probable, 2$^{\text{d}}$ is better

call T the tree of questions, $\ell_x(T)$ the nb of question to conclude x in T

$\bar{\ell}(T) = \sum_x p(x)\, \ell_x(T)$ average nb of questions asked

T^* the choice that minimizes $\bar{\ell}(T)$

claim (Shannon's ~~source~~ coding th):

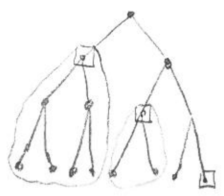
$$\boxed{H(p) \leq \bar{\ell}(T^*) < H(p) + 1}$$

⇒ $H(p)$ really quantifies the lack of information on the output of p, as the minimal nb of binary questions to ask to determine the outcome elements of the proof:

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

  ie if $|\mathcal{X}|$ grows, some $\ell_x(T)$ must be big to compensate

  proof: $\ell_{max} = \max_x \ell_x$



$2^{\ell_{max}}$ nodes at depth ℓ_{max}

each ℓ_x projects a "shadow" on $2^{\ell_{max} - \ell_x}$ nodes

shadows do not intersect $\Rightarrow \sum_{x \in \mathcal{X}} 2^{\ell_{max} - \ell_x} \leq 2^{\ell_{max}}$

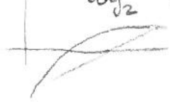
by definition $\boxed{\cdot} \Rightarrow$ answer found $\Rightarrow$ no other $\boxed{\cdot}$ below

divide by $2^{\ell_{max}}$ $\Rightarrow$ done

- Kullback Leibler divergence

  p, q two proba. laws on $\mathcal{X}$, def $D(p \| q) = \sum_{x \in \mathcal{X}} p(x) \log_2 \frac{p(x)}{q(x)}$

  properties: $*$ $D(p\|q) \geq 0$: $D = \underbrace{- \sum_{x \in \mathcal{X}} p(x) \log_2 \frac{q(x)}{p(x)}}$



$\log_2$ concave

$$\underset{\text{Jensen}}{\leq} \log_2 \left(\sum_{x \in \mathcal{X}} p(x) \frac{q(x)}{p(x)} \right) = \log_2(1) = 0$$

$*$ $D(p\|q) = 0 \Rightarrow p(x) = q(x) \quad \forall x \in \mathcal{X}$

because log strictly concave, Jensen saturated $\Rightarrow f^{\circ} = \text{cst}$

rk: with $q(x) = \frac{1}{|\mathcal{X}|}$ $\quad D(p\|q) = -H(p) + \log_2(|\mathcal{X}|) \Rightarrow H(p) \leq \log_2 |\mathcal{X}|$, equality only if $p = q$

proof of Jensen: $\mathbb{E}[f(X)] \leq f(\mathbb{E}[X])$ for f concave

$$X = \mathbb{E}[X] + (X - \mathbb{E}[X])$$

$$f(a+b) \leq f(a) + (b-a)\lambda$$

$$f(X) \leq f(\mathbb{E}[X]) + (X - \mathbb{E}[X])\lambda$$

$$\mathbb{E}[f(X)] \leq f(\mathbb{E}[X]) + 0 \cdot \lambda$$

$\lambda = f'(a)$ if f derivable, even if not derivable such a λ exists

- putting together the two

$$q(x) = \frac{1}{z} \, 2^{-\ell_x(T)} \qquad \text{for a valid tree,} \qquad z = \sum_{x \in \mathcal{X}} 2^{-\ell_x(T)}$$

$$D(p\|q) = \sum_x p(x) \log_2 \left(\frac{p(x)}{\frac{1}{z} 2^{-\ell_x(T)}} \right) = -H(p) + \log_2(z) + \bar{\ell}(T)$$

Kraft $\Rightarrow$ $z \leq 1$ $\quad \log_2(z) \leq 0$

$$\bar{\ell}(T) - H(p) = \underbrace{D(p\|q)}_{\geq 0} \underbrace{- \log_2(z)}_{\geq 0} \geq 0 \qquad \Rightarrow \qquad H(p) \leq \bar{\ell}(T) \quad \forall \text{ valid } T$$

$$\Rightarrow \quad H(p) \leq \bar{\ell}(T^*)$$

this prooves the lowerbound of Shannon's th

upperbound: one can exhibit a tree which achieves it :

* if $\{\ell_x\}$ set of integers satisfying Kraft, then $\exists T$ with $\ell_x(T) = \ell_x$:
  order the ℓ_x $\ell_1 \geq \ell_2 \geq \dots$, take the first node in lexicographic order of depth ℓ_1, then remove what is below and continue

* set $\ell_x = \left\lceil \log_2 \frac{1}{p(x)} \right\rceil \geq -\log_2 p(x)$, $\sum_x 2^{-\ell_x} \leq \sum_x p(x) = 1 \Rightarrow$ Kraft,

$$\sum_x p(x) \, \ell_x \leq \sum_x p(x) \left(\log_2 \frac{1}{p(x)} + 1 \right) = H(p) + 1$$

Huffman coding to find the optimal choice of $\{\ell_x\}$ given $p(x)$: group the two least likely symbols to make a single one, and iterate

II. C. Data compression

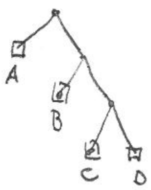
back to original problem : $x_1 x_2 \cdots x_n \longrightarrow \underbrace{0100 1 \ldots 01}$

$\in \mathcal{X}$

as short as possible
but that allows to recover

we have actually solved this problem

tree



left branch $\leftrightarrow 0$
right branch $\leftrightarrow 1$

A : 0
B : 10
C : 110
D : 111

each symbol x associated to a string $w_x(T)$ of 0,1, of length $\ell_x(T)$

$$x_1 \cdots x_n \longrightarrow w_{x_1}(T)\, w_{x_2}(T) \cdots w_{x_n}(T)$$

the string is uniquely decodable (prefix free): just follow the game.
or instantaneous, stronger than u.c., can be decoded on the fly reading only once the string

$$\underbrace{0}_{A}\,\underbrace{10}_{B}\,\underbrace{0}_{A}\,\underbrace{110}_{C}$$

if the x_i are iid with law $p(x)$, the length of the total sequence will be $n\, \bar{\ell}(T)$

$\boxed{\text{H(p) is the average number of bits (within 1) per symbol necessary to compress a sequence of symbols generated by a source of prob } p}$

rk : . use gzip to measure the entropy of English
. not iid, but short correlations
. this is for loss-less compression,
in image / sound compression some mistakes are tolerable, compromise between accuracy of reconstruction and rate of compression :
rate-distortion theory

II.D. Mutual information

useful for the following and for other lectures (Florent)

(X,Y) a pair of (a priori correlated) r.v. on $(\mathcal{X} \times \mathcal{X}')$ not necessarily the same

$p_{X,Y}(x,y) = \mathbb{P}[X=x \text{ and } Y=y]$ joint law

$$\left.\begin{aligned} p_X(x) &= \sum_{y\in\mathcal{X}'} p_{X,Y}(x,y) \\ p_Y(y) &= \sum_{x\in\mathcal{X}} p_{X,Y}(x,y) \end{aligned}\right\} \text{ marginal laws}$$

$p_{X|Y}(x|y) = \dfrac{p_{X,Y}(x,y)}{p_Y(y)}$ conditional law, $\sum_x p_{X|Y}(x|y) = 1$

if X,Y ind, $p_{X,Y}(x,y) = p_X(x)\, p_Y(y)$, $p_{X|Y}(x|y) = p_X(x)$

various entropies : $H(X,Y) = -\sum_{x,y} p_{X,Y}(x,y) \log_2 p_{X,Y}(x,y)$

$$H(X) = -\sum_x p_X(x) \log_2 p_X(x)$$

$$H(Y) = -\sum_y p_Y(y) \log_2 p_Y(y)$$

$$H(X|Y) = \sum_y p_Y(y) \left(-\sum_x p_{X|Y}(x|y) \log_2 p_{X|Y}(x|y) \right)$$

$$= -\sum_{x,y} p_{X,Y}(x,y) \log_2 p_{X|Y}(x|y)$$

mutual information between X and Y:

$$I(X;Y) = D\left(p_{X,Y} \,\|\, p_X p_Y\right) = \sum_{x,y} p_{X,Y}(x,y) \log_2 \frac{p_{X,Y}(x,y)}{p_X(x)\, p_Y(y)}$$

properties : . $I \geq 0$ (we have seen this $\forall D$)

. $I = 0 \iff p_{X,Y} = p_X p_Y \iff X$ and Y independent

. $I(X;Y) = \sum_{x,y} p_{X,Y}(x,y) \log_2 \left(\dfrac{p_{X|Y}(x|y)}{p_X(x)} \right) = H(X) - H(X|Y)$

$= H(Y) - H(Y|X)$

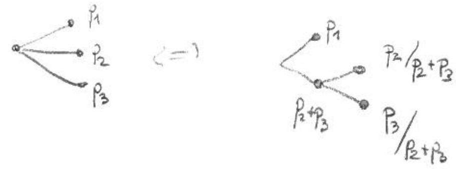
$= H(X) + H(Y) - H(X,Y)$

. hence $H(X|Y) \leq H(X)$ conditioning reduces entropy

$I(X;Y)$ measures how much you know (in bits) about one of the two r.v. if the other is revealed to you

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

additional remarks, exercices:

- proof that canonical Gibbs Boltzman is the one maximizing entropy under constraint on average energy
- proof that uniform law is the one with maximal entropy
- in Shannon's paper, proof that if one assumes
  - $H(p_1, \cdots, p_M)$ continuous
  - $H\left(\frac{1}{M}, \cdots, \frac{1}{M}\right)$ $\nearrow$ with M
  - H "additive" under decomposition 

then only possibility is $-\sum_i p_i \ln p_i$ within a multiplicative constant

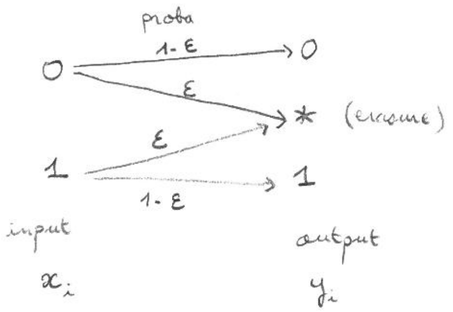
III Communication over noisy channels

III.A Definitions

$\boxed{\text{message}} \xrightarrow{\text{channel}} \boxed{\text{corrupted message}}$

wire
EM waves
Internet
...

examples . Binary Erasure Channel (BEC)



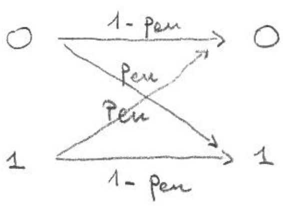
$\varepsilon = 0$ no noise

$\varepsilon = 1$ no signal at all

symetric in 0/1

if one receives 0/1 one is sure that it was the correct value

. Binary Symmetric Channel (BSC)



$p_{err} = 0$ no noise

$p_{err} = \frac{1}{2}$ no signal at all

restrict $p \le \frac{1}{2}$ by symmetry

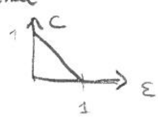
if $p_{err} > 0$ one is never sure of the result

Capacity of a channel : X the input, Y the output,

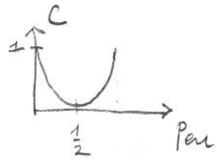
$$C = \max_{p_X} I(X;Y)$$: nb of "effective bits of information" transmitted per use of the channel

exercises: compute C for BEC and BSC, show that

. $C = 1 - \varepsilon$ for the BEC



. $C = 1 - h(p_{err})$ for the BSC, $h(p) = -p\ln p - (1-p)\ln(1-p)$



meaning will be clearer later on

proof of the capacity

. BSC $\quad p_X$ parametized by a, $p_X(0)$

$$I(X;Y) = H(Y) - \underbrace{H(Y|X)}_{h(p)}$$

$h(p)$ independently on a

$Y = 0$ with proba $a(1-p) + (1-a)p$

$\Rightarrow H(Y) = h\left(a(1-p) + (1-a)p\right)$ maximized with $a = \frac{1}{2}$, $H(Y) = h\left(\frac{1}{2}\right) = 1$

obvious a posteriori by symmetry $\quad C = 1 - h(p)$

. BEC $\quad H(Y|X) = h(\varepsilon)$ independently on a either $\{0, *\}$ or $\{1, *\}$

$$H(Y): \quad Y = \begin{cases} 0 & \text{with proba } a(1-\varepsilon) \\ 1 & (1-a)(1-\varepsilon) \\ * & \varepsilon \end{cases}$$

$$H(Y) = -\varepsilon \ln \varepsilon - a(1-\varepsilon)\ln\left(a(1-\varepsilon)\right) - (1-a)(1-\varepsilon)\ln\left((1-a)(1-\varepsilon)\right)$$

$$= -\varepsilon \ln \varepsilon - (1-\varepsilon)\ln(1-\varepsilon) - (1-\varepsilon)\, a \ln a - (1-\varepsilon)(1-a)\ln(1-a)$$

$\Rightarrow I(X;Y) = (1-\varepsilon)\, h(a)$, max in $a = \frac{1}{2}$ (again obvious) $\Rightarrow C = 1 - \varepsilon$

Encoding and decoding

noise of the channel destroys information $\Rightarrow$ fight it by transmitting more information, ie add redundancy

$$\boxed{\text{message}} \xrightarrow{\text{encoder}} \underbrace{\boxed{\text{encoded message}}}_{\text{"longer"}} \xrightarrow{\text{channel}} \boxed{\text{corrupted encoded message}} \xrightarrow{\text{decoder}} \boxed{\text{decoded message}}$$

sender and receiver agree on the encoding and decoding procedure beforehand

<u>rate of a code</u> : $\dfrac{\text{nb of bits of message}}{\text{nb of bits of encoded message}} < 1$, measure of redundancy, should be as big as possible

III.B Naive coding

simplest way to be redundant: repeat oneself

encoding : $0 \to 000$
$1 \to 111$

on the BSC, with $p < \frac{1}{2}$

decoding : $000 \to 0$

$001, 010, 100 \to 0$ majority rule (better take an odd number of repetitions to avoid it)

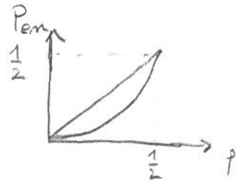
$011, 101, 110 \to 1$

$111 \to 1$

we want to transmit 0 or 1 with the same proba

without repetition, $P_{err} = p$ (just take output as a guess)

with repetition, $P_{err} = 3\,p^2(1-p) + p^3$: two or three flips induce a mistake
one can be corrected



better, but : * $R = \frac{1}{3}$

* $P_{err} > 0 \quad \forall p > 0$

to have $P_{err} \to 0$ repeat 5, 7, ... times, but then $R \to 0$
can one do better ?

III.C. Shannon channel coding theorem

quite surprisingly, answer of Shannon thm is yes:

$$\boxed{\exists \text{ codes (of growing size) with } P_{err} \to 0 \quad \text{for any } R < C}$$

* C is the ultimate limit for the rate ($P_{err} > 0$ if $R > C$, Fano inequality, intuitive given the interpretation of I)
* yet it can be approached arbitrarily close
* statement true in the thermodynamic limit

more formal definitions

message $\longrightarrow$ encoded message $\xrightarrow{\text{channel}}$ corrupted encoded message $\longrightarrow$ corrected encoded message $\longrightarrow$ corrected message

$\underline{z} = (z_1, \dots, z_L) \in \{0,1\}^L$; $\underline{x} = (x_1, \dots, x_N) \in \{0,1\}^N$; $\underline{y} = (y_1, \dots, y_N) \in \mathcal{X}_{out}^N$; $\hat{\underline{x}}(\underline{y})$; $\hat{\underline{z}}(\underline{y})$

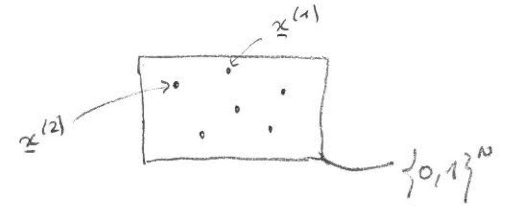
$N > L$, $R = \frac{L}{N}$

$\mathcal{X}_{out} = \{0,1\}$ BSC

or $\mathcal{X}_{out} = \{0,1,*\}$ BEC

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

ie we choose 2^L "codewords" $\underline{x}^{(\alpha)} \in \{0,1\}^N$ $\alpha = 1, 2, \dots 2^L$ to represent the possible messages



codebook: $\mathcal{C} \subset \{0,1\}^N$

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

intuitively, if $\underline{x}^{\alpha}$ is transmitted,

y has moved because of the noise.

to avoid mis-recognition the $\underline{x}^{(\alpha)}$ should be "far away" from each other $\Rightarrow$ packing problem in Hamming space, cf Cohn's lectures

but geometry in high dimension counter-intuitive, 3 points are never "aligned" in the same plane as in 2d

$\hat{\underline{x}}(\underline{y})$: estimation of the sent $\underline{x}$ given the received $\underline{y}$

then $\hat{\underline{z}}(\underline{y})$ is just $f^{-1}_{encoding}(\hat{\underline{x}}(\underline{y}))$

assuming the 2^L $\underline{z}$ are equiprobable, then the 2^L $\underline{x}^{\alpha}$ are equiprobable, we can forget about $\underline{z}$ and $\hat{\underline{z}}$: the problem is thus:

$\underline{x} \in \mathcal{C}$ u.a.r, $\to$ $\underline{y}$ corrupted $\to$ reconstruct $\underline{x}$

decoding as an inference problem (for one given codebook)

$\underline{X}$ random codeword, $P_{\underline{X}}(\underline{x}) = \frac{1}{2^L} \mathbb{1}(\underline{x} \in \mathcal{C})$ indicator function

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

with Q describing the channel, for BEC

$$Q(0|0) = 1-\varepsilon$$
$$Q(*|0) = \varepsilon$$
$$Q(1|1) = 1-\varepsilon$$
$$Q(*|1) = \varepsilon$$
$$Q(1|0) = Q(0|1) = 0$$

given $\underline{y}$ we have to guess $\underline{x}$

Bayes theorem : $$P_{\underline{X}|\underline{Y}}(\underline{x}|\underline{y}) = P_{\underline{Y}|\underline{X}}(\underline{y}|\underline{x}) \frac{P_{\underline{X}}(\underline{x})}{P_{\underline{Y}}(\underline{y})} = \frac{1}{Z(\underline{y})} \mathbb{1}(\underline{x} \in \mathcal{C}) \prod_{i=1}^{N} Q(y_i|x_i)$$

as there is noise we cannot a priori do better than assigning probabilities to the various possible input code words

$\hat{\underline{x}}(\underline{y})$? depends on the measure of error one wants to minimize

- $\hat{\underline{x}}(\underline{y}) = \underset{\underline{x}}{\operatorname{argmax}} \; P_{\underline{X}|\underline{Y}}(\underline{x}|\underline{y})$ minimizes $\mathbb{P}[\hat{\underline{X}} \neq \underline{X}]$

  block - Maximal A Posteriori (MAP) decoding
  $\Leftrightarrow$ Maximum Likelihood (ML)

- $\hat{x}_i(\underline{y}) = \underset{x_i}{\operatorname{argmax}} \; P_{X_i|\underline{Y}}(x_i|\underline{y})$ minimizes the number of badly decoded bits

  marginal law of $P_{\underline{X}|\underline{Y}}$

more on that in Krzakala's lectures

for the BEC : $\underline{y} = (0, 1, *, *, 1, 0, *, 0)$

those are sure — those are completely free

$P_{\underline{X}|\underline{Y}}(\underline{x}|\underline{y})$: uniform over those code words which agree with $\underline{y}$ on the unerased place

call $B(\underline{y}) = \{ \underline{x} \in \{0,1\}^N : \text{ if } y_i \in \{0,1\} \text{ then } x_i = y_i \}$

$$= \frac{1}{|\mathcal{C} \cap B(\underline{y})|} \mathbb{1}(\underline{x} \in \mathcal{C} \cap B(\underline{y}))$$

Shannon Random Code Ensemble

reminiscent of REM (Biroli's lectures), also in the treatment

$\mathcal{C} = \{ \underline{x}^{(\alpha)} \}_{\alpha = 1, \dots, 2^L}$ obtained by taking $x_i^{(\alpha)} = \begin{cases} 0 & \text{proba } 1/2 \\ 1 & \text{"} \end{cases}$ independently $\forall i, \alpha$

very simple, seems crazy, does not ensure that the $\underline{x}^{(\alpha)}$ are all different, but in the limit $N, L \to \infty$ very few collisions

on the BEC : suppose w.l.o.g that $\underline{x}^{(1)}$ has been transmitted

$N_*(\underline{y})$: nb of erased bits in $\underline{y}$, r.v. $\mathrm{Bin}(N, \varepsilon)$ $\simeq N\varepsilon$ in the thermodynamic limit

how many codewords $\neq \underline{x}^{(1)}$ in $B(\underline{y})$?

r.v. $\mathcal{N} = \mathrm{Bin}\left(2^L - 1, \left(\frac{1}{2}\right)^{N - N_*(\underline{y})} \right)$

↖ proba that $\underline{x}^{(\alpha \neq 1)}$ agrees with $\underline{x}^{(1)}$ on the non erased bits

$$\mathbb{E}[\mathcal{N}] \simeq 2^{L - N + \varepsilon N} = 2^{N(R - 1 + \varepsilon)} \underset{N \to \infty}{\longrightarrow} 0 \quad \text{if} \quad R < 1 - \varepsilon = C$$

$\Rightarrow$ as in the REM, or first moment method $\mathcal{N} = 0$ w.h.p

$\Rightarrow$ one can recover $\underline{x}^{(1)}$ as the only codeword compatible with the received bits in $\underline{y}$

$\Rightarrow$ proves Shannon's channel coding theorem for the BEC,

one can decode with vanishing error probability with rates up to capacity

proof for the BSC as homework, text and solution (in french) on the webpage,

⚠ notations inverted

very simple random code achieves capacity

BUT : . needs exponentially large (in N) memory to store the codewords

. needs exponential time to decode

$\Rightarrow$ needs to add structure to the codebook, can this be done and still achieves capacity?

to be seen in the following

rk : • here we have average (with respect / codewords and to the code) error probability $\to 0$, can be boosted to maximal (over the codewords) error probability $\to 0$, hence $\exists$ one code with max error proba $\to 0$ by expurgating the worst half codewords, does not change the rate

• converse with Fano, $P_{err} > 0$ if $R > C$, actually $P_{err} \to 1$ exp. fast in N

IV Low Density Parity Check Codes (LDPC)

IV.A. Linear codes

need to add some structure to the codebook $\mathcal{C} = \{\underline{x}^{(\alpha)}\} \subset \{0,1\}^N$ to make encoding and decoding easier

$\{0,1\}^N$ is a linear space over $\{0,1\} = \mathbb{Z}_2$, with

$0+0=0$, $0+1=1+0=1$, $1+1=0$ (addition mod 2)

$0 \cdot 0 = 0$, $0 \cdot 1 = 0$, $1 \cdot 0 = 0$, $1 \cdot 1 = 1$

$\mathcal{C}$ is a linear code if it is a linear subspace of $\{0,1\}^N$

ie $\begin{cases} \underline{0} \in \mathcal{C} \text{ and} \\ \underline{x}, \underline{y} \in \mathcal{C} \end{cases} \Rightarrow \underline{x}+\underline{y} = (x_1+y_1, \dots, x_N+y_N) \in \mathcal{C}$ (mod 2)

would be enough, $\underline{x}+\underline{x} = \underline{0} \Rightarrow$ imply first

linear algebra on $\{0,1\}$ instead of $\mathbb{R}$, works partly the same

- $\dim \mathcal{C}$: nb of linearly independent code words $\Rightarrow |\mathcal{C}| = 2^{\dim \mathcal{C}}$
- $\mathcal{C}$ can be specified as $\operatorname{Ker} H = \{ \underline{x} \in \{0,1\}^N ,\ H\underline{x} = 0 \}$, H a $M \times N$ matrix on $\mathbb{Z}_2$ $\Rightarrow$ 0,1 entries

or as $\operatorname{Im} G = \{ \underline{u}\, G ,\ u \in \{0,1\}^{N-M} \}$, G a $N-M \times N$ matrix

H parity check matrix, one line of $H\underline{x}=0$ of the form $x_{12} + x_{42} + x_{45} = 0$

parity of the number of 1's must be even

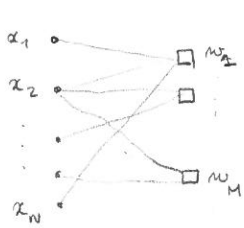
if these equations are linearly independent, $|\mathcal{C}| = 2^{N-M}$

$\dim \operatorname{Ker} H = N - \operatorname{rk} H$ always, independence $\Leftrightarrow$ full rank

rate of the code $R = \frac{N-M}{N} = 1 - \frac{M}{N}$ (assuming equations independent, in particular $M<N$)

G generator matrix, allows to encode the codewords very easily

Tanner / factor graph representation of parity checks:



$$w_a(\underline{x}) = \mathbb{1}\left(\sum_{i=1}^N H_{ai}\, x_i = 0 \right)$$

edge between a and i $\Leftrightarrow$ $H_{ai} = 1$

$\partial a = \{ i : H_{ai} = 1 \}$, $\partial i = \{ a : H_{ai} = 1 \}$ neighborhoods on the graphs

all codewords are equivalent in a linear code, and $(0,\dots,0)$ is always a codeword

rk : addition mod 2 of $\{0,1\}$ $\Leftrightarrow$ eXclusive OR of $\{\text{True}, \text{False}\}$ $\rightarrow$ known as a CSP as XORSAT mentionné par Guis

- $\sigma_i = (-1)^{x_i} = \begin{cases} 1 & x_i = 0 \\ -1 & x_i = 1 \end{cases}$ $\quad \sum_i x_i = 0 \mod 2 \Leftrightarrow \prod_i \sigma_i = 1$: also p-spin interactions

$\Rightarrow$ lectures of Biroli, Ricci-Tersenghi, Moore...

IV. B. Definition of the simplest ensemble

given the success of Shannon's random code, try a random H (Gallager 62)

but "low density" : a few 1's by row and column of H $\rightarrow$ few operations, faster computationally

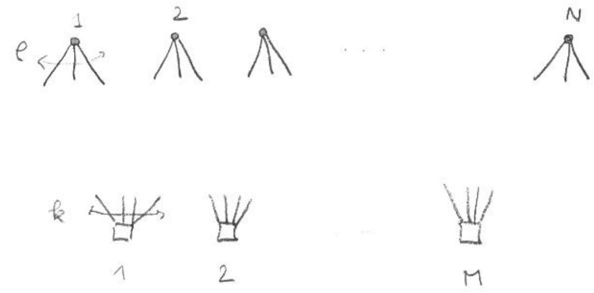
take H such that : $|\partial i| = \ell \quad \forall i$, ie ℓ 1's in each column

$|\partial a| = k \quad \forall a$ $\qquad k$ row

with ℓ, k finite in the thermodynamic limit

$\Rightarrow \quad N\ell = Mk$, total number ~~of~~ 1's in the matrix $\qquad R = 1 - \frac{\ell}{k}$

how to do this in practice : $\qquad \ell < k$



$\ell = 3$

$k = 4$

here

random matching (permutation) of these $N\ell = Mk$ half edges

there can be pb if

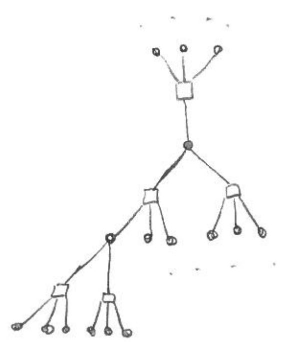
can eliminate even nb of parallel edges

$x_i + x_i = 0 \mod 2 \quad \forall x_i$

doesnot really matter, will be rare in the thermodynamic limit

if $M < N$ the equations are linearly independent (or subextensive nb are dependent)

<u>crucial property</u> : locally tree-like in the thermodynamic limit



no short loops (with high probability)

intuitive explanation : exploring from one vertex, choose finite nb of neighbors out of an extensive one, proba to choose twice in the same finite set is $O(1/N)$

⚠ not the tree of the guessing game

IV. C Analysis on the BEC

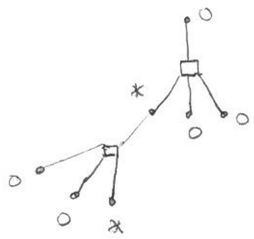
assume codeword $(0,\dots,0)$ has been transmitted over BEC(ϵ)

→ because all codewords are equivalent, but we must preted not to know that it is $(0,\dots,0)$

received $\underline{y} = (0, *, 0, 0, \dots, *)$

sender and receiver have agreed on an LDPC

what can the receiver do ?



the first * can be set to 0 with the first parity check

then the second * as well

- in all checks where there is a single * → set it to 0
- continue iteratively

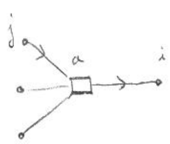
when it stops,
- either no * remains → perfect decoding
- or at least two * in the checks that remain

stopping set : subset of the var such that each check contains either 0 or ≥ 2 var of the stopping set

polynomial time algorithm

up to which value of ϵ will the first situation occurs ?

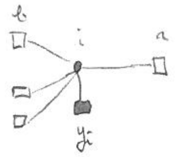
reformulation of the algorithm as message passing (cf Moore's lectures) $u_{a\to i}, h_{i\to a} \in \{0, *\}$



$$u_{a\to i} = \begin{cases} 0 & \text{if } h_{j\to a} = 0 \quad \forall j \in \partial a \setminus i \\ * & \text{otherwise} \end{cases}$$

"I'm sure you are 0"

"I can't say"



$$h_{i\to a} = \begin{cases} * & \text{if } y_i = * \text{ and } u_{b\to i} = * \quad \forall b \in \partial i \setminus a \\ 0 & \text{otherwise} \end{cases}$$

"I'm sure I'm a 0, either because I've received my bit, or deduced it from one neighboring check"

$$h_i = \begin{cases} * & \text{if } y_i = * \text{ and } u_{a\to i} = * \quad \forall a \in \partial i \\ 0 & \text{otherwise} \end{cases}$$

final estimate

in discrete time $u_{a\to i}^{(t=0)} = *$ for all $a \to i$

$$h_{i\to a}^{(t=0)} = f\left(\{u^{(t=0)}, \underline{y}\}\right)$$

$$u_{a\to i}^{(t=1)} = f\left(\{h^{(t=0)}\}\right)$$

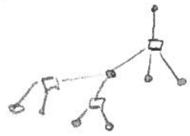
$h_i^{(t)}$: estimate from the $u^{(t)}$

monotonicity, messages switch from $*$ to 0, at the end same final state for the h_i than the original decoder, for any update schedule

probabilistic analysis

$\eta^{(t)} = \text{Proba}\left(u^{(t)}_{a\to i} = 0 \right)$ with respect to
- choice of the H
- choice of the pattern of erasures (channel)
- choice of an edge uniformly at random

$\xi^{(t)} = \text{Proba}\left(h^{(t)}_{i\to a} = 0 \right)$ idem



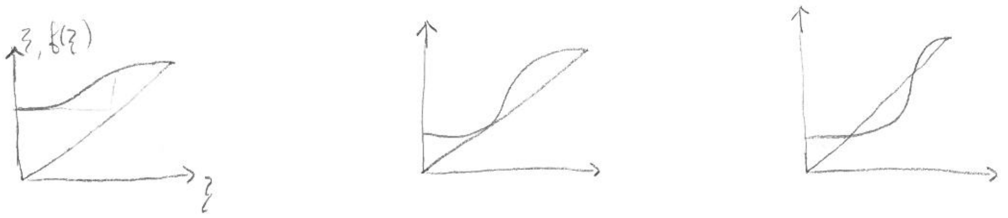
$$\left\{ \begin{array}{l} \eta^{(t=0)} = 0 \\ \eta^{(t+1)} = \left(\xi^{(t)}\right)^{k-1} \\ \xi^{(t)} = 1 - \varepsilon \left(1 - \eta^{(t)}\right)^{\ell-1} \end{array} \right.$$

all inputs must be 0, iid because of the tree structure

all must be $*$ for h to be $*$

$$\xi^{(t)} = f\left(\xi^{(t-1)}\right), \qquad f(\xi) = 1 - \varepsilon\left(1 - \xi^{k-1}\right)^{\ell-1}$$

properties of f: $f(0) = 1-\varepsilon$, $f(1) = 1$, $f \nearrow$ — if $k \geq 3$, $\ell \geq 3$, zero derivatives in 0 and 1

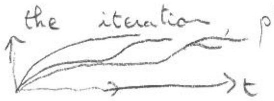


ε small — $\varepsilon < \varepsilon_{BP}$

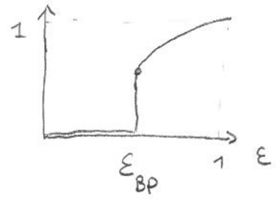
$\varepsilon = \varepsilon_{BP}$

ε big — $\varepsilon > \varepsilon_{BP}$

plot of the iteration, plateau in $\xi^{(t)}$, exponent $1/2$, similarity with $C(t)$



$$\mathbb{P}\left[h_i^{(t=\infty)} = * \right] = \varepsilon \left(1 - \xi^{(\infty)\,k-1}\right)^{\ell}$$



for $\varepsilon < \varepsilon_{BP}$, perfect decoding, in linear time, for a positive rate

$\Rightarrow$ non trivial results

numerical values

not yet defined

ℓ	k	ε_{BP}	ε_{MAP}	R	ε_{Sh}
3	4	0,647	0,746	1/4 = 0,25	0,75
3	5	0,518	0,591	2/5 = 0,4	0,6
3	6	0,429	0,488	1/2 = 0,5	0,5
4	6	0,506	0,665	1/3 = 0,333	0,666

$\varepsilon_{Sh} = 1 - R$ the maximal level of noise a code of this rate could correct

$\varepsilon_{BP} < \varepsilon_{Sh}$ is this because of the code ?

or because of the algorithm to decode ?

in other words, for $\varepsilon > \varepsilon_{BP}$, once BP has infered all possible implications,

one has a linear system of equations on N' variables, M' equations,

has this system . a single solution ? then perfect decoding would still be possible by exhaustive algorithm (or Gauss elimination here)

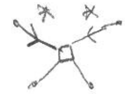
. or more than one solution, then perfect decoding is not possible, whatever the algorithm

$$N' = N\,\varepsilon\left(1-\xi^{k-1}\right)^{\ell}$$

with $\xi = \xi^{\infty}$, the number of undecoded variables

$$M' = M\left(1-\xi^{k} - k\,\xi^{k-1}(1-\xi)\right)$$

at least two in the incoming messages



one needs to work with messages and not h_i to have independence

assuming the equations are independent,

$$\text{nb of solutions} = 2^{N'-M'} = 2^{N\phi}$$

$$\text{with } \phi(\varepsilon,\xi) = \varepsilon\left(1-\xi^{k-1}\right)^{\ell} - \frac{\ell}{k}\left(1-\xi^{k} - k\,\xi^{k-1}(1-\xi)\right)$$

$$\frac{d}{d\xi}\phi = -\varepsilon\,\ell(k-1)\,\xi^{k-2}\left(1-\xi^{k-1}\right)^{\ell-1} + \ell\left(\xi^{k-1} + (k-1)\xi^{k-2} - k\,\xi^{k-1}\right)$$

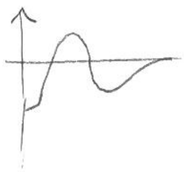
$$= \ell(k-1)\,\xi^{k-2}\left[-\varepsilon\left(1-\xi^{k-1}\right)^{\ell-1} + 1 - \xi\right]$$

$= 0$ when ξ is a fixed point of the iterations

$\varepsilon < \varepsilon_{BP}$ $\qquad$ $\varepsilon = \varepsilon_{BP}$ inflexion point $\qquad$ $\varepsilon \in [\varepsilon_{BP}, \varepsilon_{MAP}]$ $\qquad$ $\varepsilon = \varepsilon_{MAP}$

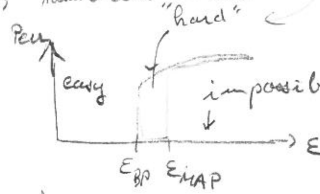
like Franz-Parisi potential (with − sign, $1-\xi \leftrightarrow q$)

plots shown by Florent but not exactly the same phases, $m=0$ vs $m=1$ in terms of magnetization



$\varepsilon > \varepsilon_{MAP}$

$\Rightarrow$ numerical values "hard" ← forgetting Gaussian elimination



to do perfect decoding

computational gap, as in many inference pb (Florent)

$\simeq$ like T_d / T_K, but not completely

IV. D. More generic LDPC ensembles

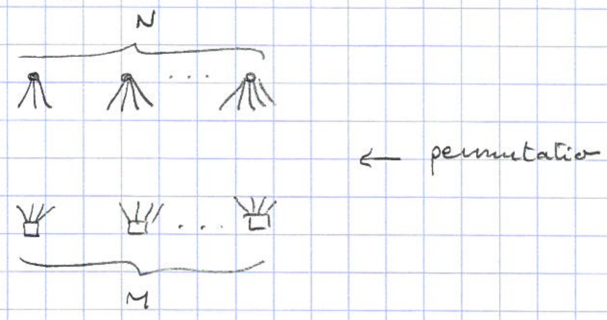
instead of fixed degrees (ℓ, k) distribution of degrees

variables, checks

λ_ℓ : fraction of variables that have degree ℓ

ρ_k : ——— checks ——— k

H uniform under this constraint



$\langle \ell \rangle = \sum_\ell \ell \lambda_\ell$, $\langle k \rangle = \sum_k k \rho_k$ $\qquad N \langle \ell \rangle = M \langle k \rangle$ to match

still locally tree like

if one draws at random an edge, ends up on ∘ with degree $\ell+1$

with proba $\tilde{\lambda}_\ell = \dfrac{(\ell+1)\lambda_{\ell+1}}{\langle \ell \rangle}$ $\qquad$ on □ with degree $k+1$

with proba $\tilde{\rho}_k = \dfrac{(k+1)\rho_{k+1}}{\langle k \rangle}$

define $\begin{cases} \lambda(x) = \sum_\ell \lambda_\ell x^\ell & \rho(x) = \sum_k \rho_k x^k \\ \tilde{\lambda}(x) = \sum_\ell \tilde{\lambda}_\ell x^\ell , & \tilde{\rho}(x) = \sum_k \tilde{\rho}_k x^k \end{cases}$ $\qquad \tilde{\lambda}(x) = \dfrac{\lambda'(x)}{\lambda'(1)}$

then, <u>exercise</u> : $\begin{cases} \eta^{(t+1)} = \tilde{\rho}\left(\xi^{(t)}\right) \\ \xi^{(t)} = 1 - \varepsilon\, \tilde{\lambda}\left(1 - \eta^{(t)}\right) \end{cases}$ $\qquad$ same reasoning as before, just add the proba of finding a vertex of a certain degree

$$\phi(\varepsilon, \xi) = \varepsilon\, \lambda\left(1 - \tilde{\rho}(\xi)\right) - \frac{\langle \ell \rangle}{\langle k \rangle}\left(1 - \rho(\xi) - \langle k \rangle (1-\xi)\, \tilde{\rho}(\xi)\right)$$

generalizes the case with fixed k, ℓ

one can find choices of $\{\lambda_\ell, \rho_k\}$ such that ε_{BP} is arbitrarily close to ε_{Sh}

seems great but

linear decoding up to capacity

not obvious that they are the best in practice, depends on the finite N behavior, and prefactor in the linear complexity diverges when $\varepsilon_{BP} \to \varepsilon_{Sh}$ (max. degree $\nearrow$)

V Conclusions and perspectives

what we have done is a very specific (and simple) example of the reasoning of the cavity method : Let's abstract the main points, besides the IT context :

. variables $\underline{x} = (x_1, \dots, x_N) \in \mathcal{X}^N$ (here the encoded message to be recovered)

. quenched randomness $\underline{J}$ (here the graph / parity check matrix + realization of the noise)

defining a factor graph
with M interactions

• probability law $\mu(\underline{x};\underline{J}) = \frac{1}{Z(\underline{J})} \prod_{i=1}^{N} w_i(x_i;\underline{J}) \prod_{a=1}^{M} w_a(\underline{x}_{\partial a};\underline{J})$

here : information received from the channel ; constraints on the codewords

goals of the cavity method :

compute • $\phi = \lim_{N\to\infty} \frac{1}{N} \mathbb{E}\left[\ln Z(\underline{J})\right]$ quenched free entropy density (here ≈ conditional entropy)

• marginal laws $\mu(x_i;\underline{J})$ "magnetizations"

strategy : . tree like factor graph

. trees exactly solved by message passing (here the messages were 0/1/* in general non trivial proba distribution on $\mathcal{X}$)

. long loops → "irrelevant" → RS cavity method
↘ induce long range correlations → RSB

⚠ LDPC = planted models, on the Nishimori line, don't take as completely generic the phenomenon / transitions found here

⟹ cf Federico's lectures

Sourlas 89
Richardson - Urbanke
Luby - Mitzenmacher - Shokrollahi - Spielman } 2001
Kabashima - Saad ~ 2000
Montanari ~ 2000