ICFP M2 - Statistical physics 2 – TD n° 6
Communication over a noisy channel

Grégory Schehr, Guilhem Semerjian

February 2020

A person wants to send a message to another one; the message is represented as a string of N bits, $\underline{x} = (x_1, \dots, x_N) \in \{0,1\}^N$, and the communication between the two persons occurs through a noisy channel. Each bit sent through the channel is transmitted correctly with probability $1-p$, and flipped to the opposite value with probability p (with $p \leq 1/2$). If $\underline{x}$ is transmitted directly through the channel errors are unavoidable as soon as $p > 0$. To correct these errors the sender and the receiver agree beforehand to use some error-correcting code. To each of the $M = 2^N$ possible values of $\underline{x}$ they associate a codeword $\underline{z}^{(\alpha)}$, with $\alpha = 1, 2, \dots, M$, where the codewords are strings of $L > N$ bits, $\underline{z}^{(\alpha)} = (z_1^{(\alpha)}, \dots, z_L^{(\alpha)}) \in \{0,1\}^L$. We denote $\mathcal{C} = \{\underline{z}^{(1)}, \dots, \underline{z}^{(M)}\}$ the set of all codewords. When the sender wants to transmit the message $\underline{x}$ he sends instead through the channel the corresponding codeword $\underline{z}^{(\alpha)}$; the receiver gets at the output of the channel a corrupted version $\underline{z}' \in \{0,1\}^L$ of the codeword. We shall see that if the code is well constructed and the noise level p not too high the redundancy introduced with the $L - N$ supplementary bits allows to decode the message with a probability of error that goes to zero in the $N \to \infty$ limit.

1. Compute the probability that, given that the receiver gets the string $\underline{z}'$, the sender had tried to communicate the codeword $\underline{z} \in \mathcal{C}$. You should use the Bayes theorem, assume that all the M codewords are a priori equiprobable, and denote $d(\underline{z}, \underline{z}')$ the Hamming distance between two binary strings (i.e. the number of positions where they differ).

2. In the following we assume that the receiver decodes the channel output by selecting the codeword closest (in Hamming distance) to $\underline{z}'$; justify this choice with your answer to the previous question.

Following the original proposal of Shannon we consider that the set of codewords that the sender and receiver agree to use is generated randomly : the $M \times L$ bits $z_i^{(\alpha)}$ are chosen, independently of each other, to be 0 or 1 with probability 1/2.

Without loss of generality we assume in the following that the transmitted codeword is $\underline{z}^{(1)}$.

3. We denote $\mathcal{N}$ the number of codewords among $\underline{z}^{(2)}, \dots, \underline{z}^{(M)}$ that are at Hamming distance $D \in [0, L]$ from $\underline{z}^{(1)}$.

   (a) Show that $\mathcal{N}$ has a binomial distribution, and give its parameters.

   (b) What is the average value of $\mathcal{N}$?

   (c) We consider the thermodynamic limit $N, L \to \infty$, with the rate of the code $R = N/L$ fixed, and $D = L\,\delta$ with $\delta \in [0,1]$ fixed. Show that at the exponential order

   $$\mathbb{E}[\mathcal{N}] \doteq 2^{L f(R,\delta)} \ , \qquad \text{with} \quad f(R,\delta) = R - 1 - \delta \log_2 \delta - (1-\delta)\log_2(1-\delta) \ .$$

   (d) Draw the function f as a function of δ, and conclude on the existence of a minimal (intensive) Hamming distance $\delta_*(R)$ between a given codeword and the closest distinct codeword in a typical realization of the code (the reasoning here is similar to the one followed for the Random Energy Model).

4. What is the law of the Hamming distance $d(\underline{z}^{(1)}, \underline{z}')$ between the transmitted codeword and the output of the channel ? Conclude that if $p < \delta_*(R)/2$ the codeword sent can typically be recovered without error by the receiver.

5. Actually this is possible for $p < \delta_*(R)$. To realize why, consider now $\mathcal{N}$ as the number of codewords among $\underline{z}^{(2)}, \ldots, \underline{z}^{(M)}$ that are at Hamming distance $D \in [0, L]$ from $\underline{z}'$, the output of the channel. Is there something to modify in your computation of the statistics of $\mathcal{N}$?

It turns out that, despite their simplicity, these random codes are the best possible from the point of view of the noise level they can correct. Shannon's channel coding theorem asserts indeed that for any code with rate R, the probability of error cannot go to zero in the large size limit if $p > \delta_*(R)$. However these random codes have very bad algorithmic performances: encoding and decoding a message of length N takes a time which is exponentially large in N. In recent years new random constructions of codes have been discovered, that combine the noise performance of Shannon construction with much faster algorithms ; these ensembles are tightly related to mean-field spin glasses on random graphs.