

ICFP M2 - STATISTICAL PHYSICS 2 – TD n° 5

Random XORSAT problems

Guilhem Semerjian

February 2019

We shall consider in this problem random systems of linear equations, also known as XORSAT problems, denoted F . We recall their definition given during the lectures :

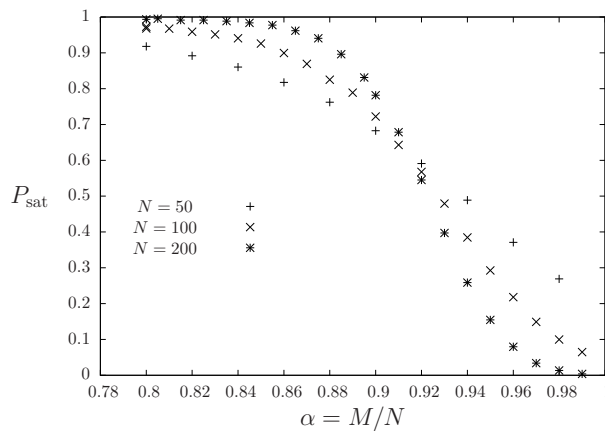
- the degrees of freedom are N Boolean variables, $\underline{x} = (x_1, \dots, x_N) \in \{0, 1\}^N$
- they have to obey M linear constraints of the form

$$x_{i_a^1} + x_{i_a^2} + \dots + x_{i_a^k} = y_a \pmod{2}, \quad (1)$$

where $a = 1, \dots, M$ indexes the various equations, $k \geq 3$ is an integer defining the number of variables involved in each equation, $\langle i_a^1, \dots, i_a^k \rangle$ is a k -uplet of distinct indices in $\{1, \dots, N\}$, and $y_a \in \{0, 1\}$ fixes the right hand side of the equation.

- such a formula is said to be satisfiable if there is a configuration \underline{x} that verifies all the equations simultaneously, unsatisfiable otherwise.
- a random ensemble of formulas is defined very easily by generating the M equations independently, choosing for each of them a k -uplet $\langle i_a^1, \dots, i_a^k \rangle$ uniformly at random among the $\binom{N}{k}$ possible ones, and $y_a = 0$ or 1 with probability $1/2$.

Using the Gaussian elimination algorithm one can determine whether a given formula is satisfiable or not in polynomial time. Repeating this process a large number of times one can easily obtain a numerical estimate of the probability $P_{\text{sat}}(\alpha, N)$ that a random formula F with N variables and $M = \alpha N$ equations is satisfiable :



These curves, obtained for $k = 3$, suggest that a phase transition occurs in the thermodynamic limit ($N, M \rightarrow \infty$ with $\alpha = M/N$ fixed) for α around 0.92. Indeed, there exists a threshold α_{sat} (that depends on k) such that

$$\lim_{N \rightarrow \infty} P_{\text{sat}}(\alpha, N) = \begin{cases} 1 & \text{if } \alpha < \alpha_{\text{sat}} \\ 0 & \text{if } \alpha > \alpha_{\text{sat}} \end{cases}. \quad (2)$$

1 Bounds on α_{sat}

We recall a result obtained in TD2 : for a random variable Z that takes non-negative integer values,

$$\frac{\mathbb{E}[Z]^2}{\mathbb{E}[Z^2]} \leq \mathbb{P}[Z > 0] \leq \mathbb{E}[Z], \quad (3)$$

these two inequalities being called the second and first moment method, respectively.

We shall use these inequalities with Z the number of solutions of a random XORSAT formula with N variables and M equations constructed as above.

1. Compute $\mathbb{E}[Z]$, and deduce that $\alpha_{\text{sat}} \leq 1$.
2. Show that

$$\mathbb{E}[Z^2] = 2^N \sum_{D=0}^N \binom{N}{D} \left(\frac{1}{2} \sum_{\substack{l=0 \\ l \text{ even}}}^k \frac{\binom{D}{l} \binom{N-D}{k-l}}{\binom{N}{k}} \right)^M. \quad (4)$$

3. Study the asymptotic behavior of this expression in the thermodynamic limit, and conclude that

$$\lim_{N \rightarrow \infty} \frac{1}{N} \ln \left(\frac{\mathbb{E}[Z]^2}{\mathbb{E}[Z^2]} \right) = \inf_{d \in [0,1]} g(\alpha, d), \quad \text{with } g(\alpha, d) = \ln 2 + d \ln d + (1-d) \ln(1-d) - \alpha \ln(1 + (1-2d)^k).$$

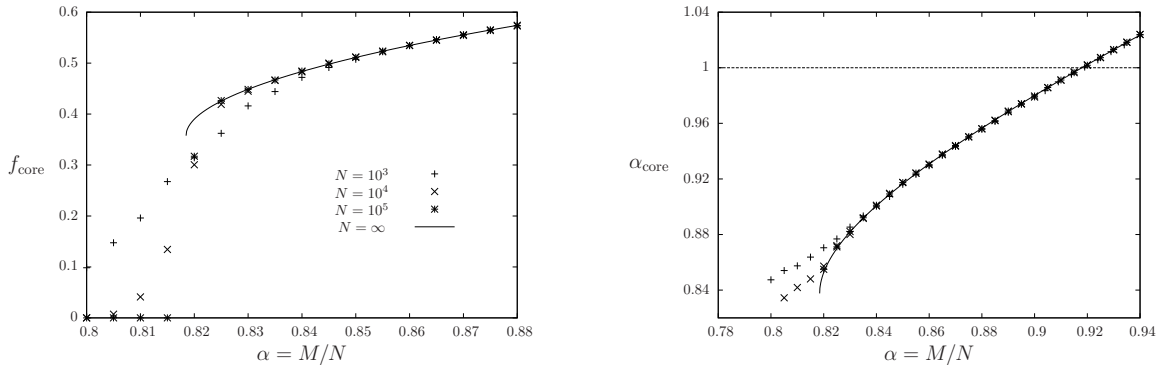
4. Draw the shape of the function g as a function of d for increasing values of α . Conclude that there exists a value $\alpha_{\text{lb}} > 0$ (equal to 0.889 for $k = 3$) such that for $\alpha < \alpha_{\text{lb}}$, the first term in (3) is not exponentially small. A more detailed analysis of (4) shows that in this case it actually goes to 1. Conclude that $\alpha_{\text{lb}} \leq \alpha_{\text{sat}} \leq 1$.

2 Leaf removal procedure

The bounds on α_{sat} obtained above are not tight (i.e. $\alpha_{\text{lb}} < 1$) because of the potentially huge fluctuations of Z , that can cause its average $\mathbb{E}[Z]$ to be much larger than its typical value. These fluctuations can be reduced by concentrating on a well-chosen subformula, as explained now.

1. Suppose that F contains a leaf, i.e. a variable that appears in a single equation, and denote F' the system of equations obtained by removing this equation. Show that F is satisfiable if and only if F' is satisfiable.

This leaf removal procedure can be iterated as long as leaves are present. Two cases can occur : either the formula is completely emptied by this procedure, or there remains a non-trivial subset of F , called its core, in which every variable appears in at least two equations. We call N_{core} and M_{core} the number of variables and equations of the core formula, and display on the curves below the fraction $f_{\text{core}} = N_{\text{core}}/N$ of variables in the core and the density $\alpha_{\text{core}} = M_{\text{core}}/N_{\text{core}}$ of equations it contains.



These curves demonstrate a core percolation transition at $\alpha_d = 0.818$ (for $k = 3$), and show that the density α_{core} crosses 1 at $\alpha_* = 0.918$ (for $k = 3$).

2. A calculation (not required here) shows that

$$f_{\text{core}} = 1 - e^{-\alpha k \phi^{k-1}} - \alpha k \phi^{k-1} e^{-\alpha k \phi^{k-1}}, \quad \frac{1}{N} M_{\text{core}} = \alpha \phi^k, \quad (5)$$

where $\phi = \phi(\alpha, k)$ is the largest solution in $[0, 1]$ of the equation

$$\phi = 1 - e^{-\alpha k \phi^{k-1}}. \quad (6)$$

Study graphically this equation, show that for $k \geq 3$ the transition at α_d is discontinuous, and study the behavior of ϕ for $\alpha \rightarrow \alpha_d^+$.

3. Explain why α_* is an improved upperbound on α_{sat} . It turns out that the fluctuations of the core are much smaller than that of the full formula, hence actually $\alpha_{\text{sat}} = \alpha_*$.