3

SOLUTIONS TO EXERCISES

3.1 Exercise 1: Detailed study of 1-XORSAT

Figure 3.1(top) shows the probability P_{SAT} that a randomly extracted 1-XORSAT formula is satisfiable as a function of the ratio α, and for sizes N ranging from 100 to 1000. We see that P_{SAT} is a decreasing function of α and N.

Consider the subformula made of the n_i equations with first member equal to x_i. This formula is always satisfiable if $n_i = 0$ or $n_i = 1$. If $n_i \geq 2$ the formula is satisfiable if and only if all second members are equal (to 0, or to 1), an event with probability $(\frac{1}{2})^{n_i-1}$ decreasing exponentially with the number of equations. Hence we have to consider the following variant of the celebrated Birthday problem[16]. Consider a year with a number N of days, how should scale the number M of students in a class to be sure that no two students have the same birthday date?

$$\bar{p} = \prod_{i=0}^{M-1} \left(1 - \frac{i}{N}\right) = \exp\left(-\frac{M(M-1)}{2N} + O(M^3/N^2)\right) . \qquad (3.1)$$

Hence we expect a cross-over from large to small $\bar{p}$ when M crosses the scaling regime $\sqrt{N}$. Going back to the 1-XORSAT model we expect P_{SAT} to have a non zero limit value when the number of equations and variables are both sent to infinity at a fixed ratio $y = M/\sqrt{N}$. In other words, random 1-XORSAT formulas with N variables, M equations or with, say, $100 \times N$ variables, $10 \times M$ equations should have roughly the same probabilities of being satisfiable. To check this hypothesis we replot the data in Figure 3.1 after multiplication of the abscissa of each point by $\sqrt{N}$ (to keep y fixed instead of α). The outcome is shown in the bottom panel of Figure 3.1. Data obtained for various sizes nicely collapse on a single limit curve function of y.

The calculation of this limit function, usually called scaling function, is done hereafter in the fixed-probability 1-XORSAT model where the number of equations is a Poisson variable of mean value $\bar{M} = y\sqrt{N}$. We will discuss the equivalence between the fixed-probability and the fixed-size ensembles later. In the fixed-probability ensemble the numbers n_i of occurence of each variable x_i are

[16]The Birthday problem is a classical elementary probability problem: given a class with M students, what is the probability that at least two of them have the same birthday date? The answer for $M = 25$ is $p \simeq 57\%$, while a much lower value is expected on intuitive grounds when M is much smaller than the number $N = 365$ of days in a year.

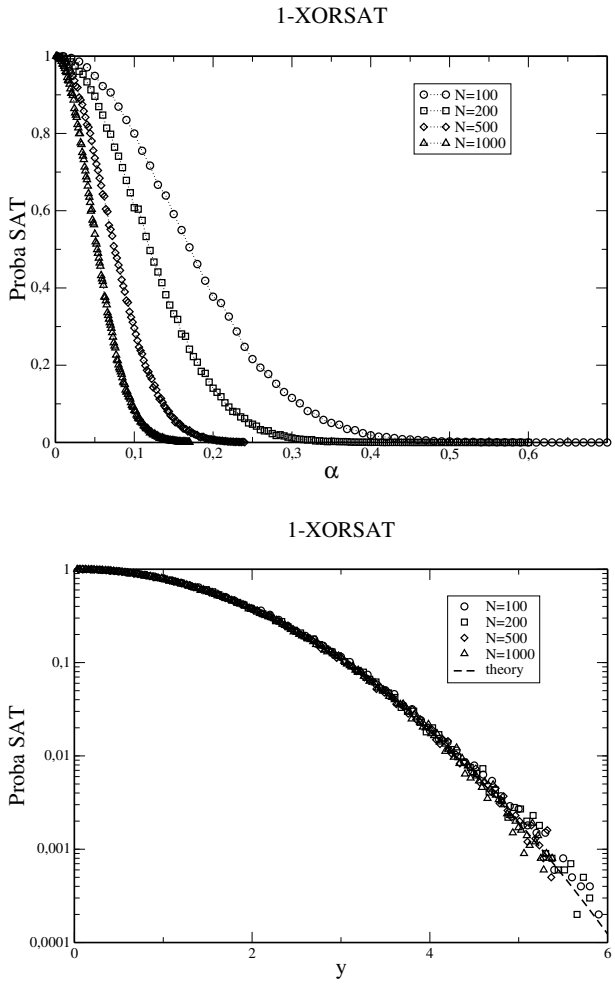

FIG. 3.1. Top: Probability that a random 1-XORSAT formula is satisfiable as a function of the ratio α of equations per variable, and for various sizes N. Bottom: same data as in the left panel after the horizontal rescaling $\alpha \to \alpha \times \sqrt{N} = y$; note the use of a log scale for the vertical axis. The dashed line shows the scaling function $\Phi_1(y)$ (3.3).

independent Poisson variables with average value $\bar{M}/N = y/\sqrt{N}$. Therefore the probability of satisfaction is

$$P^p_{SAT}(N, \alpha = \frac{y}{\sqrt{N}}) = \left[e^{-y/\sqrt{N}} \left(1 + \sum_{n \geq 1} \frac{(y/\sqrt{N})^n}{n!} \left(\frac{1}{2} \right)^{n-1} \right) \right]^N$$

$$= \left[2e^{-y/(2\sqrt{N})} - e^{-y/\sqrt{N}}\right]^N \ , \tag{3.2}$$

where the p subscript denotes the use of the fixed-probability ensemble. We obtain the desired scaling function

$$\Phi_1(y) \equiv \lim_{N\to\infty} \ln P^p_{SAT}(N, \alpha = \frac{y}{\sqrt{N}}) = -\frac{y^2}{4} \ , \tag{3.3}$$

in excellent agreement with the rescaled data of Figure 3.1 (bottom) [**?**].

For finite but large N there is a tiny probability that a randomly extracted formula is actually satisifiable even when $\alpha > 0$. A natural question is to characterize the 'rate' at which P_{SAT} tends to zero as N increases (at fixed α). Answering to such questions is the very scope of large deviation theory. Looking for events with very small probabilities is not only interesting from an academic point of view, but can also be crucial in practical applications.

Figure 3.2 shows minus the logarithm of P_{SAT}, divided by N, as a function of the ratio α and for various sizes N. Once again the data corresponding to different sizes collapse on a single curve, meaning that

$$P_{SAT}(N, \alpha) = e^{-N\ \omega_1(\alpha) + o(N)} \ . \tag{3.4}$$

Decay exponent ω_1 is called rate function in probability theory. We can derive its value in the fixed-probability ensemble from (3.2) with $y = \alpha \times \sqrt{N}$, with the immediate result

$$\omega_1^p(\alpha) = \alpha - \ln\left(2\ e^{\alpha/2} - 1\right) \ . \tag{3.5}$$

The agreement with numerics is very good for small ratios, but deteriorates as α increases. The reason is simple. In the fixed-probability ensemble the number M of equations is not fixed but may fluctuate around the average value $\bar{M} = \alpha N$. The ratio $\tilde{\alpha} = M/N$, is with high probability equal to α, but large deviations ($\tilde{\alpha} \neq \alpha$) are possible and described by the rate function[17],

$$\Omega(\tilde{\alpha}|\alpha) = \tilde{\alpha} - \alpha - \alpha\ \ln(\alpha/\tilde{\alpha}) \ . \tag{3.6}$$

However the probability that a random 1-XORSAT formula with M equations is satisfiable is also exponentially small in N, with a rate function $\omega_1(\alpha)$ increasing with α. Thus, in the fixed-probability ensemble, a trade-off is found between ratios $\tilde{\alpha}$ close to α (formulas likely to be generated) and close to 0 (formulas likely to be satisfiable). As a result the fixed-probability rate function is

[17] M obeys a Poisson law with parameter $\bar{M}$. Using Stirling formula,

$$e^{-\bar{M}}\frac{\bar{M}^M}{M!} \simeq e^{-\alpha N}(\tilde{\alpha}N)^{\alpha N}\sqrt{2\pi N}\left(\frac{e}{\alpha N}\right)^{\alpha N} = e^{-N\,\Omega(\tilde{\alpha}|\alpha)+o(N)} \ ,$$

where Ω is defined in (3.6).

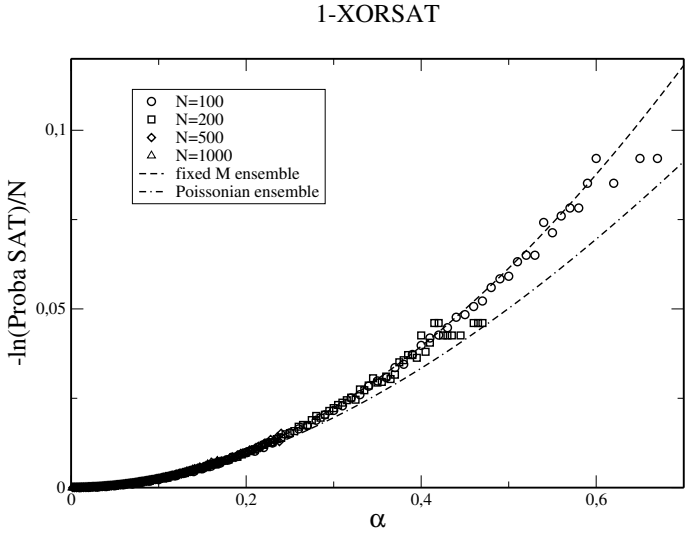

FIG. 3.2. Same data as Figure 3.1 with: logarithmic scale on the vertical axis, and rescaling by $-1/N$. The scaling functions ω_1 (3.7) and ω_1^p (3.5) for, respectively, the fixed-size and fixed-probability ensembles are shown.

$$\omega_1^p(\alpha) = \min_{\tilde{\alpha}} \left[\omega_1(\tilde{\alpha}) + \Omega(\tilde{\alpha}|\alpha)\right] , \tag{3.7}$$

and is smaller than $\omega_1(\alpha)$. It is an easy check that the optimal ratio $\tilde{\alpha}^* = \alpha/(2 - e^{-\alpha/2}) < \alpha$ as expected. Inverting (3.7) we deduce the rate function ω_1 in the fixed-size ensemble, in excellent agreement with numerics (Figure 3.2). This example underlines that thermodynamically equivalent ensembles have to be considered with care as far as rare events are concerned.

Remark that, when $\alpha \to 0$, $\tilde{\alpha} = \alpha + O(\alpha^2)$, and $\omega_1^p(\alpha) = \omega_1(\alpha) + O(\alpha^3)$. This common value coincides with the scaling function $-\Phi_1(\alpha)$ (3.3). This identity is expected on general basis, and justifies the agreement between the fixed-probability scaling function and the numerics based on the fixed-size ensemble (Figure 3.1, right).

3.2 Exercise 2: dynamics of the UC heuristics

Let α_0 denote the equation per variable ratio of the 3-XORSAT instance to be solved. We call $E_j(T)$ the number of j–equations (including j variables) after T variables have been assigned by the solving procedure. T will be called hereafter 'time', not to be confused with the computational effort. At time $T = 0$ we have $E_3(0) = \alpha_0 N$, $E_2(0) = E_1(0) = 0$. Assume that the variable x assigned at time T is chosen from a single-variable clause, that is, independently of the j-equation content. Call $n_j(T)$ the number of occurrences of x in j-equations ($j = 2, 3$). The evolution equations for the populations of 2-,3-equations read

$$E_3(T+1) = E_3(T) - n_3(T) \ , \quad E_2(T+1) = E_2(T) - n_2(T) + n_3(T) \ . \quad (3.8)$$

Flows n_2, n_3 are of course random variables that depend on the instance under consideration at time T, and on the choice of variable done by UC. What are their distributions? At time T there remain $N-T$ untouched variables; x appears in any of the $E_j(T)$ j-equation with probability $p_j = \frac{j}{N-T}$, independently of the other equations. In the large N limit and at fixed fraction of assigned variables, $t = \frac{T}{N}$, the binomial distribution converges to a Poisson law with mean

$$\langle n_j \rangle_T = \frac{j\, e_j}{1-t} \qquad \text{where} \qquad e_j = \frac{E_j(T)}{N} \quad (3.9)$$

is the density of j-equations at time T. The key remark is that, when $N \to \infty$, e_j is a slowly varying and non stochastic quantity and is a function of the fraction $t = \frac{T}{N}$ rather than T itself. Let us iterate (3.8) between times $T_0 = t\,N$ and $T_0 + \Delta T$ where $1 \ll \Delta T \ll N$ e.g. $\Delta T = O(\sqrt{N})$. Then the change ΔE_3 in the number of 3-equations is (minus) the sum of the stochastic variables $n_j(T)$ for $T = T_0, T_0+1, \ldots, T_0+\Delta T$. As these variables are uncorrelated Poisson variables with $O(1)$ mean (3.9) ΔE_3 will be of the order of ΔT, and the change in the density e_3 will be of order of $\Delta T/N \to 0$. Applying central limit theorem $\Delta E_3/\Delta T$ will be almost surely equal to $-\langle n_3 \rangle_t$ given by (3.9) and with the equation density measured at reduced time t. The argument can be extended to 2-equations, and we conclude that e_2, e_3 are deterministic (self-averaging) quantities obeying the two coupled differential equations

$$\frac{de_3}{dt}(t) = -\frac{3\, e_3}{1-t} \quad , \qquad \frac{de_3}{dt}(t) = \frac{3\, e_3}{1-t} - \frac{2\, e_2}{1-t} \ . \quad (3.10)$$

Those equations, together with the initial condition $e_3(0) = \alpha_0$, $e_2(0) = 0$ can be easily solved,

$$e_3(t) = \alpha_0 (1-t)^3 \quad , \qquad e_2(t) = 3\,\alpha_0\, t\, (1-t)^2 \ . \quad (3.11)$$

To sum up, the dynamical evolution of the equation populations may be seen as a slow and deterministic evolution of the equation densities to which are superimposed fast, small fluctuations. The distribution of the fluctuations adiabatically follows the slow trajectory. This scenario is pictured in Figure 3.3.

The trajectories we have derived in the previous Section are correct provided no contradiction emerges. But contradictions may happen as soon as there are $E_1 = 2$ unit-equations, and are all the more likely than E_1 is large. Actually the set of 1-equations form a 1-XORSAT instance which is unsatisfiable with a finite probability as soon as E_1 is of the order of $\sqrt{N}$ from the results of Exercise 1. Assume now that $E_1(T) \ll N$ after T variables have been assigned, what is the probability ρ_T that no contradiction emerges when the T^{th} variable is assigned by UC? This probability is clearly one when $E_1 = 0$. When $E_1 \geq 1$ we pick up a 1-equation, say, $x_6 = 1$, and wonder whether the opposite 1-equation,

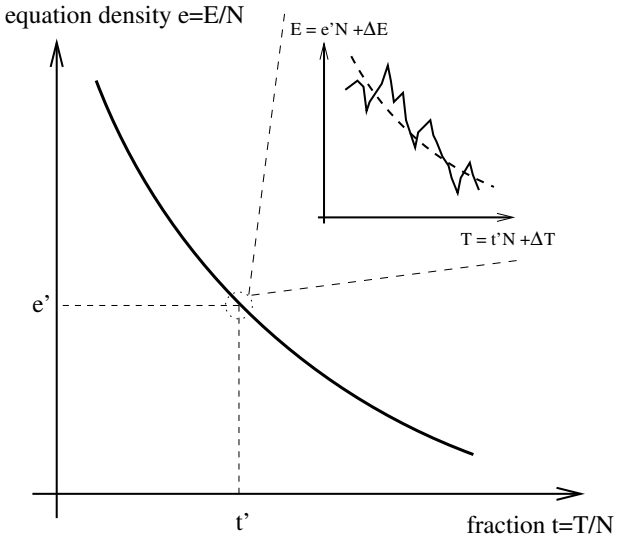

FIG. 3.3. Deterministic versus stochastic dynamics of the equation population E as a function of the number of steps T of the algorithm. On the slow time scale (fraction $t = T/N$) the density $e = E/N$ of (2- or 3-) equations varies smoothly according to a deterministic law. Blowing up of the dynamics around some point t', e' shows the existence of small and fast fluctuations around this trajectory. Fluctuations are stochastic but their probability distribution depends upon the slow variables t', e' only.

$x_6 = 0$, is present among the $(E_1 - 1)$ 1-equations left. As equations are uniformly distributed over the set of $N - T$ untouched variables

$$\rho_T = \left(1 - \frac{1}{2(N-T)}\right)^{\max(E_1(T)-1,0)} . \tag{3.12}$$

The presence of the max in the above equation ensures it remains correct even in the absence of unit-equations ($E_1 = 0$). $E_1(T)$ is a stochastic variable. However from the decoupling between fast and slow time scales sketched in Figure 3.3 the probability distribution of $E_1(T)$ depends only on the slow time scale t. Let us call $\mu(E_1; t)$ this probability. Multiplying (3.12) over the times $T = 0$ to $T = N - 1$ we deduce the probability that DPLL has successfully found a solution without ever backtracking,

$$\rho_{success} = \exp\left(-\int_0^1 \frac{dt}{2(1-t)} \sum_{E_1 \geq 1} \mu(E_1; t)\ (E_1 - 1)\right) \tag{3.13}$$

in the large N limit.

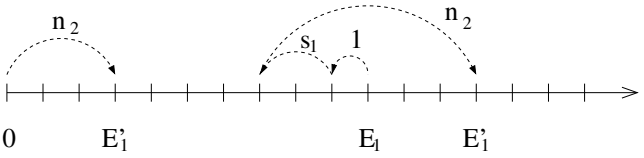

FIG. 3.4. Evolution of the number E_1 of 1-equations as one more variable is assigned. n_2 denotes the number of 2-equations reduced to 1-equations, s_1 the number of 1-equations satisfied. If $E_1 \geq 1$ a variable is fixed through unit-propagation: E_1 decreases by one plus s_1, and increases by n_2. In the absence of unit-equation ($E_1 = 0$) the number of 1-equations after the assignment is simply $E_1' = n_2$.

We are left with the calculation of μ. Figure 3.4 sketches the stochastic evolution of the number E_1 during one step. The number of 1-equations produced from 2-equations, n_2, is a Poisson variable with average value, from (3.11),

$$d(t) = \frac{2\,e_2(t)}{1-t} = 6\,\alpha_0\,t(1-t) \tag{3.14}$$

when $N \to \infty$. The number of satisfied 1-equations, s_1, is negligible as long as E_1 remains bounded. The probability that the number of 1-equations goes from E_1 to E_1' when $T \to T+1$ defines the entry of the transition matrix

$$M(E_1', E_1; t) = \sum_{n_2 \geq 0} e^{-d(t)} \frac{d(t)^{n_2}}{n_2!} \delta_{E_1' - (E_1 + n_2 - \delta_{E_1})} \ . \tag{3.15}$$

from which a master equation for the probability of E_1 at time T may be written. On time scales $1 \ll \Delta T \ll N$ this master equation converges to the equilibrium distribution μ, conveniently expressed in terms of the generating function

$$G(x;t) = \sum_{E_1 \geq 0} \mu(E_1;t)\ x^{E_1} = \frac{(1-d(t))(x-1)}{x\ e^{d(t)\ (1-x)} - 1} \quad . \tag{3.16}$$

The above is a sensible result for $d(t) \leq 1$ but does not make sense when $d(t) > 1$ since a probability cannot be negative! The reason is that we have derived (3.16) under the implicit condition that no contradiction was encountered. This assumption cannot hold when the average rate of 1-equation production, $d(t)$, is larger that one, the rate at which 1-equations are satisfed by unit-propagation. From (3.14) we see, when $\alpha > \alpha_E = \frac{2}{3}$, the trajectory would cross the

$$\alpha_D(p) = \frac{1}{2(1-p)} \tag{3.17}$$

on which $d = 1$ for some time $t_D < 1$. A contradiction is very likely to emerge before the crossing.

When $\alpha < \alpha_E$ d remains smaller than unity at any time. In this regime the probability of success reads, using (3.13) and (3.16),

$$\rho_{success} = \exp\left(\frac{3\alpha}{4} - \frac{1}{2}\sqrt{\frac{3\alpha}{2-3\alpha}}\ \tanh^{-1}\left[\sqrt{\frac{3\alpha}{2-3\alpha}}\right]\right) \ . \tag{3.18}$$

$\rho_{success}$ is a decreasing function of the ratio α, down from unity for $\alpha = 0$ to zero for $\alpha = \alpha_E$. In can be shown that, right at α_E, $\rho_{success} \sim \exp(-Cst \times N^{\frac{1}{6}})$ decreases as a stretched exponential of the size. The value of the exponent, and its robustness against the splitting heuristics are explained in Deroulers, Monasson, Critical behaviour of combinatorial search algorithm and the unit-clause universality class, Europhysics Letters 68, 153 (2004).