Catching the k-NAESAT threshold

Konstantinos Panagiotou

Séminaire de Combinatoire Philippe Flajolet
Institut Henri Poincaré, Séance du 27 septembre 2012

1 Problem statement

This presentation is focused on the random k-NAESAT problem, which is one of the standard benchmark problems in the theory of random Constraint Satisfaction Problems (CSPs). The input to the problem consists of a Boolean formula in k-conjunctive normal form (k-CNF). An assignment of values to the variables is called *not-all-equal satisfying* (NAESAT) if it is satisfying and there is no clause in which all literals are satisfied. Note that for any given k-clause there are $2^k - 2$ assignments of the literals that satisfy the clause.

In the random setting, the problem is as follows: Suppose that $x_1, \ldots, x_n$ are the n random variables, and let $m = \lceil cn \rceil$ for some real $c > 0$. Let $F(n, m)$ denote a k-CNF formula with m clauses, where each clause is drawn uniformly at random from the set of all possible clauses. The central question one can ask in this context is for which c is $F(n, m)$ NAE-satisfiable *with high probability* (whp)? This is called the *threshold* of random k-NAESAT.

2 Summary of previous work and contribution of this presentation

For the cases $k = 1, 2$ the threshold is well understood. More generally, Friedgut [4] showed that there is a sharp threshold *sequence* $c_k(n)$ such that if $c < c_k(n)$ then $F(n, cn)$ is satisfiable whp whereas if $c > c_k(n)$ it is unsatisfiable whp. Achlioptas and Moore [2] gave upper and lower bounds for $c_k(n)$ of the form $2^{k-1} \ln 2 - \frac{1+\ln 2}{2} < c_k(n) < 2^{k-1} \ln 2 - \frac{\ln 2}{2}$. The lower bound for $c_k(n)$ was improved by Coja-Oghlan and Zdeborová [5] who showed that $2^{k-1} \ln 2 - \ln 2 < c_k(n)$. This left an additive gap of $\frac{1}{2} \ln 2 \approx 0.347$ which this work closes. Namely, we showed that $c_k(n)$ is equal to

$$2^{k-1} \ln 2 - \left(\frac{\ln 2}{2} + \frac{1}{4} \right) + \varepsilon_k$$

where $\varepsilon_k < 2^{-k}$.

This improvement, albeit modest at first sight, is conceptually significant for two reasons. First, we obtain (virtually) matching upper and lower bounds for the first time in a random CSP of this type. Second, we devise a rigorous method for understanding what happens at the so-called *condensation phase*, which occurs shortly before the threshold phase.

The k-NAESAT problem belongs in the class of random CSP problems (along with many other problems such as random k-SAT, k-coloring random graphs and 2-coloring random k-uniform hypergraphs). For random CSPs, statistical physicists have developed sophisticated but non-rigorous techniques, which, nevertheless, have provided a detailed picture about the structural properties and have helped raise many conjectures.

3 Shattering and condensation

What is the evolution of the solution space of k-NAESAT? To answer this, consider the following graph G_c. The vertices of G_c are all NAE-satisfying assignments of $F_{n,cn}$; moreover, there is an edge between two solutions if and only if their Hamming distance is small, namely $o(n)$. Clearly, if $c = 0$ then V contains all 2^n assignments, and G_0 is connected. On the other extreme, if $c > c_k(n)$, then G_c is empty whp.

There are two significant phases in the evolution of G_c, as c increases. The first phase, namely the *shattering* phase, occurs at about c_k/k. Here, G_c contains exponentially many clusters (hence the term "shattering"), with each cluster containing exponentially many assignments. Furthermore, the pairwise distance between clusters is roughly $n/2$. This phase is well-understood, thanks to previous work of Achlioptas and Ricci-Tersenghi [3] and Achlioptas and Coja-Oghlan [1]. In contrast, the *condensation* phase occurs at $c = c_k - \left(\frac{\ln 2}{2} - \frac{1}{4}\right)$, as demonstrated in previous work by Coja-Oghlan and Zdeborová [5].

This latter phase introduces significant difficulties in the analysis. Consider the following experiment. Suppose we choose two solutions uniformly at random from the set of NAE-satisfying assignments. In the shattering phase, we expect the Hamming distance between the solutions to be roughly $n/2$. However, in the condensation phase, this distance is very small, namely $o_k(1)$. Intuitively, this means that there are heavy correlations between solutions, which explains why previously known methods break down at the condensation threshold.

4 Outline of our approach

In order to tame the difficulties observed at the condensation phase, we need to address two separate problems. The first has to do with counting *atypical* assignments, namely assignments contained in small clusters. Physicists have already provided evidence that in almost all assignments in a cluster, most variables are *frozen*, i.e, they take the same value. The problem is that there is simply no way to tell whether a given variable is frozen: deciding this is NP-hard in the worst case. Instead, we are going to work with a simple parameter that turns out to be a good substitute for the frozen variables. To this end, observe that if a variable x is frozen, then there is at least one clause C such that if we assigned x the opposite value, then C would be violated. We call a variable *blocked* if it is contained in such a clause. We were able to show that most blocked variables are frozen, and thus suffices to count NAE-satisfying assignments with sufficiently many blocked variables.

To understand why we need to fix further parameters of the formula, let us define the degree d_x of a variable x as the number of times that x occurs in the random formula F. Let $d = (d_x)_x \in V$ be the degree sequence of F. It is well known that in the "plain" random formula the degree of each variable is asymptotically Poisson with mean km/n. On the other hand, if we condition on some specific satisfying assignment that has "too many" blocked variables, then the degrees are not asymptotically Poisson anymore. Indeed, the degree d_x is the sum of the number s_x of clauses that x supports, and the number d'_x of times that x appears otherwise. While d'_x is asymptotically Poisson with mean smaller than km/n as the non-critical clauses do not affect the number of blocked variables at all, s_x is not, since it corresponds to an atypical outcome of a random experiment. The precise distribution of s_x is quite non-trivial, but it is not difficult to verify that s_x does not have a Poisson distribution.

References

[1] D. Achlioptas, A. Coja-Oghlan: Algorithmic barriers from phase transitions. Proc. 49th FOCS (2008) 793–802.

[2] D. Achlioptas, C. Moore: Random k-SAT: two moments suffice to cross a sharp threshold. SIAM Journal on Computing **36** (2006) 740–762.

[3] D. Achlioptas, F. Ricci-Tersenghi: On the solution space geometry of random constraint satisfaction problems. Proc. 38th STOC (2006) 130–139.

[4] E. Friedgut: Hunting for sharp thresholds. Random Struct. Algorithms **26** (2005) 37–51

[5] A. Coja-Oghlan, L. Zdeborová: The condensation transition in random hypergraph 2-coloring. Proc. 23rd SODA (2012), to appear.