The solution space structure of random constraint satisfaction problems with growing domains

Wei Xu\(^1\), Pan Zhang\(^{2,3}\), Tian Liu\(^{4,6}\) and Fuzhou Gong\(^5\)

\(^1\) LMIB and School of Mathematics and Systems Science, Beihang University, Beijing 100191, People’s Republic of China
\(^2\) Santa Fe Institute, Santa Fe, NM 87501, USA
\(^3\) State Key Laboratory of Theoretical Physics, Institute of Theoretical Physics, Chinese Academy of Sciences, Beijing 100190, People’s Republic of China
\(^4\) Key Laboratory of High Confidence Software Technologies, Ministry of Education, School of EECS, Peking University, Beijing 100871, People’s Republic of China
\(^5\) Institute of Applied Mathematics, Academy of Mathematics and System Science, Chinese Academy of Sciences, Beijing 100190, People’s Republic of China

E-mail: lt@pku.edu.cn

Received 12 September 2015
Accepted for publication 26 October 2015
Published 10 December 2015

Online at stacks.iop.org/JSTAT/2015/P12006
doi:10.1088/1742-5468/2015/12/P12006

Abstract. In this paper we study the solution space structure of model RB, a standard prototype of the constraint satisfaction problem (CSP), with growing domains. Using the \textit{first moment method} and the \textit{second moment method}, we rigorously show that in the satisfiable phase close to the satisfiability transition, solutions are clustered into an exponential number of well-separated clusters, with each cluster containing a sub-exponential number of solutions. As a consequence, the system has a clustering (dynamical) transition but no condensation transition. This picture of the phase diagram is different to other classic random CSPs that possess a fixed domain size, such as the \(K\)-satisfiability (\(K\)-SAT) problem and the graph colouring problem, where a condensation transition exists and is distinctly different to the satisfiability transition. Our result verifies some non-rigorous results obtained using the cavity method from spin glass theory.

Keywords: disordered systems (theory), phase transformations (theory)
1. Introduction

Constraint satisfaction problems (CSPs) are defined as a set of discrete variables whose assignments must satisfy a collection of constraints. A CSP instance is said to be satisfiable if there exists a solution, i.e. a configuration to all of the variables that satisfies all of the constraints. The core question to CSPs is to answer whether a given instance is satisfiable. CSPs have been studied extensively in mathematics and in computer science, and they have played an important role in computational complexity theory. Most of the interesting CSPs, such as Boolean $K$-satisfiability and the graph colouring problem, belong to the class of NP-complete: in the worst case scenario, the time required to decide whether there exists a solution increases very quickly with the size of the CSP.

In recent years, a lot of attention has been paid to the average-case complexity of CSPs, which studies the computational complexity of solving random ensembles of CSPs. From the viewpoint of statistical physics, finding the solutions to a given CSP amounts to finding the ground-state configurations of spin glasses at zero temperature, where the energy represents the number of violated constraints. Most of the interesting CSPs display a spin glass behaviour at the thermodynamic limit (when the number of variables $N \to \infty$, and the number of constraints $M \to \infty$), where they encounter a set of phase transitions when the constraint density $c = \frac{M}{N}$ increases. The first transition to have drawn a lot of interest is the satisfiability transition $c_s$ [1–3], where the probability of a random instance being satisfiable changes sharply from 1 to 0. When considering the satisfiable phase (a parameter regime where w.h.p.$^7$ random instances are satisfiable), studies use the cavity method [4, 5] from spin glass theory to illustrate that

$^7$‘with high probability’ (w.h.p.) means that the probability of some event tends to 1, as $N \to \infty$. 

doi:10.1088/1742-5468/2015/12/P12006
the solution space of CSPs is highly structured: as \( c \) increases, the system undergoes a clustering transition, a condensation transition and a satisfiability transition, respectively \([6–8]\). All of these transitions are connected to the fact that the solutions are clustered. This clustering phenomenon is believed to affect the performance of solution-finding algorithms, and is thought to be responsible for the hardness of CSPs \([6]\).

Besides heuristic analyses that use the cavity method, rigorous mathematical studies have also made a lot of progress towards understanding the satisfiability transition and the clustering of solutions in CSPs. Some models, such as the \( K \)-XORSAT, and the \( K \)-SAT problem with growing clause length, have been proven to have a satisfiability transition. Other models, such as \( K \)-SAT \((K \geq 8)\), \( K \)-colouring and hypergraph 2-colouring \([9–11]\) problems have been proved to have a clustering phase. Moreover, the hypergraph 2-colouring problem has been proven to have a condensation phase \([12]\).

In this paper we study model RB \([13]\), which is a prototype CSP model with growing domains revised from the famous CSP Model B \([14]\). The main difference between the model RB and classic CSPs such as the \( K \)-SAT problem is that the number of states (called domain size here) that one variable can take is an increasing function of the number of variables. This property is probably one of the reasons that make the satisfiability threshold rigorously solvable \([13]\), and make the clustering of solutions provable, as we will show in this paper. It has been shown that random instances of model RB are hard to solve in the parameter regime close to the satisfiability transition \([15–18]\), so benchmarks based on model RB (more information at www.nlsde.buaa.edu.cn/kexu/) have been widely used in algorithmic research and in various kinds of algorithm competitions (e.g. CSP, SAT and MaxSAT) in recent years.

The statistical mechanics of model RB has been extensively studied using the cavity method \([18, 19]\). It has been shown that the replica symmetry solution is always stable in the satisfiable phase, which suggests that a condensation transition does not exist in this model. Here we use rigorous methods, namely the first and the second moment methods \([10, 20]\), to show that in the satisfiable phase close to the satisfiability transition, solutions are always grouped into an exponential number of clusters, where each cluster contains a sub-exponential number of solutions. As such, we rigorously demonstrate that the system indeed has no condensation transition.

The main contributions of this paper are twofold:

- From the mathematical point of view, we give a rigorous analysis of the geometry of solution clusters in the model RB problem.
- From the statistical physics point of view, we show that there is no condensation transition in this problem. Thus, replica symmetry results including Bethe entropy and the marginals given by the cavity method and the associated belief propagation algorithm are asymptotically exact.

The rest of the paper is organised as follows. Section 2 includes definitions of model RB and brief descriptions of previously obtained results regarding phase transitions. Section 3 contains our main results, which include rigorous analyses of the clustering of solutions, as well as the number and diameters of the clusters. We conclude our work in section 4.
2. Model RB and phase transitions

Random CSP models provide relatively ‘unbiased’ samples for testing algorithms, which allow researchers to design better algorithms and heuristics. They also provide insights into computational complexity theory. It is well known that standard random CSP models (such as model B) suffer from (trivial) insolubility as the problem size increases. Thus, models with varying scales of parameters have been proposed to overcome this deficiency [21–25]. Model RB is one of them, which has been designed to have a growing domain size. This property helps describe many practical problems such as the $N$-queens problem, the Latin square problem, Sudoku and the Golomb ruler problem.

Here is the definition of model RB. An instance of model RB contains $N$ variables, each of which takes values from its domain $D = \{1, 2, \ldots, d_N\}$, with $d_N = N^\alpha$. Note that the domain size $|D|$ grows polynomially with the system size $N$, which is the main difference between model RB and classic CSPs such as the $K$-SAT problem. There are $M = rN\ln N$ constraints in one instance, where each constraint involves $k$ ($k \geq 2$) different variables that are randomly (and uniformly by default) chosen from all of the variables. The total number of assignments of variables involved in a constraint is $d_N^k$. For each constraint (denoted by $a$), we randomly pick $p d_N^k$ different assignments from the totally $d_N^k$ assignments to form an incompatible-set $Q_a$. In other words, the constraint $a$ is satisfied by the assignment $\{\sigma_a\} = \{\sigma_{a1}, \sigma_{a2}, \ldots, \sigma_{ak}\}$ if $\{\sigma_a \not\in Q_a\}$.

Given parameters $(N, k, r, \alpha, p)$, an instance of model RB is generated as follows:

1. Select (with repetition) $rN\ln N$ random constraints, each of which is formed by randomly selecting (without repetition) $k$ variables.

2. For each constraint, form an incompatible set by randomly selecting (without repetition) $pN^\alpha k$ elements.

Here we consider

$$p < 1 - \frac{1}{k},$$

in order to exclude the situation that there are too few satisfiable assignments in each constraint.

Given an instance of model RB, the task is to find a solution, i.e. a configuration that satisfies all of the constraints simultaneously. It is easy to see that the total number of configurations is $N^{\alpha N}$, each of which satisfies all of the constraints with probability $(1 - p)^{rN\ln N}$. If we use $X$ to denote the number of solutions in a given instance, its expectation over all possible instances can be written as

$$\mathbb{E}(X) = N^{\alpha N}(1 - p)^{rN\ln N}. \quad (2)$$

Let

$$r^* = -\frac{\alpha}{\ln(1 - p)}, \quad (3)$$
we can see that when \( r > r^* \), the expectation \( \mathbb{E}(X) \) is nearly 0 for large \( N \). Using Markov’s inequality,

\[
P(X > 0) \leq \mathbb{E}(X),
\]

we know that \( \mathbb{E}(X) \) gives an upper bound for the probability of a formula being satisfiable. Thus, when \( r > r^* \) w.h.p. there is no solution in a random instance of model RB. When \( r < r^* \), the expectation of the number of solutions is larger than 0. But these solutions may be distributed non-uniformly, that is, some instances may contain an exponential amount of solutions, while other instances may contain no solution at all.

Fortunately, in model RB it has been shown [13] that when \( r < r^* \), \( \mathbb{E}(x) \) tends to be the square root of the second moment of the number of solutions, hence the solutions are indeed distributed uniformly. More precisely, as \( N \to \infty \), using Cauchy’s inequality, when \( r < r^* \) we have

\[
P(X > 0) \geq \frac{\mathbb{E}^2(X)}{\mathbb{E}(X^2)} \to 1.
\]

In other words, the satisfiability transition happens at \( r^* \):

\[
\lim_{n \to \infty} \Pr(X > 0) = 1 \text{ when } r < r^*, \quad \lim_{n \to \infty} \Pr(X > 0) = 0 \text{ when } r > r^*.
\]

However, even in the satisfiable phase close to the satisfiability transition, where we are sure that there are solutions, it is still difficult to find a solution in a random instance. Indeed, there have been many efforts to design efficient algorithms that work in this regime. So far, our understanding of this algorithmic hardness is based on the clustering of solutions in the satisfiable regime close to the satisfiability transition. In statistical physics, the methods that we can use to describe the solution space structure are borrowed from the cavity method in spin glass theory. From the statistical physics point of view, CSP problems are nothing but spin glass models at zero temperature, where the energy of the system is defined as the number of violated constraints in a given CSP. Thus, finding a solution is equivalent to finding a configuration \( \{\sigma\} = \{\sigma_i| i = 1, \ldots, N\} \) that has zero energy. More precisely, one can define a Gibbs measure

\[
P(\{\sigma\}) = \frac{1}{Z} e^{-\beta \sum_{i=1}^N E_i(\{\sigma_i\})},
\]

where \( Z \) is the partition function, \( E_i(\{\sigma_i\}) = 0 \) if \( \{\sigma_i\} \not\in Q_i \) and is 1 otherwise. By taking \( \beta \to \infty \), and using e.g. the cavity method, one can study the properties of this Gibbs distribution which reflect the structure of the solution space [6]. The properties include whether the ground-state energy is 0, whether the Gibbs distribution is extremal, whether the replica symmetry is broken etc. Previous studies [6–8] have shown that similar pictures of solution space structure and phase transition exist in many interesting constraint satisfaction problems. When the number of constraints is small, replica symmetry holds and the Gibbs measure is extremal. While the number of constraints (or edges in the factor graph) increases, the system undergoes clustering, condensation and satisfiability transitions, respectively. At the clustering transition (also called the dynamical transition), the set of solutions begins to split into an exponential number...
of pure states, where replica symmetry holds in each pure state. At the condensation transition, the cluster size becomes inhomogeneous in such a way that a finite number of clusters contain almost all of the solutions. Beyond the satisfiability transition, both clusters and solutions cease to exist. In some CSPs, such as the $K$-SAT problem with $K = 3$, and some combinatorial optimisation problems such as the independent set problem [26, 27] with a low average degree, the clustering transition and the condensation transition are identical. While for some other problems such as the $K$-SAT problem with $K \geq 4$, and the graph colouring problem with more than 3 colours, the condensation transition is distinctly different to the clustering transition, and there is a stable one-step-replica-symmetry-breaking (1RSB) phase.

Studies based on the replica symmetry cavity method and its associated belief propagation equations have been applied to model RB in [19], and Bethe entropy $S_{\text{Bethe}}$ (the leading order of the logarithm of the number of solutions) has been calculated for single instances. There are two interesting observations in [19]. First, BP equations always converge on a single instance. This means that in the satisfiable phase BP is always marginally stable, indicating that the replica symmetry solution is always locally stable in the satisfiable phase. Second, the Bethe entropy agrees very well with the logarithm of the first moment of the number of solutions (annealed entropy), i.e. $S_{\text{Bethe}} = \ln \mathbb{E}(X)$. These two phenomena suggest that the belief propagation algorithm may give asymptotically correct marginals and free energy, and a condensation transition does not exist. However, decimation algorithms based on BP marginals may still fail to find a solution, because BP marginals may be incorrect after some variables are assigned, when the solutions are strongly correlated.

### 3. Solution space structure of model RB

Heuristic analyses of the solution space structure using the cavity method and replica symmetry breaking are based on the concept of pure states. These analyses assume an extremal Gibbs measure and an exponential growth of both the number of clusters and the number of solutions in each cluster. Following [28], in this article we make use of a more concrete definition of clusters using the Hamming distance. The Hamming distance between two arbitrary configurations $x$, $y$, denoted as $d(x, y)$, is the number of variables that take different values in $x$ and $y$. Denote the set of all solutions in an instance by $S$. We define the diameter of a set of solutions $X \subseteq S$ as the maximum Hamming distance between any two elements of $X$. The distance between two sets $X$, $Y \subseteq S$, is the minimum Hamming distance between any $x \in X$ and any $y \in Y$. We define a cluster as a connected component of $S$, where any two solutions $x, y \in S$ are considered adjacent if they are at a Hamming distance of one (or a finite integer $q$; this does not affect the conclusion). We further define a region as the union of some non-empty clusters.

#### 3.1. Clustering of solutions

Our analysis is based on the number of solution pairs $Z(x)$ at a Hamming distance $xN$, with $0 < x < 1$. Again, since it is hard to exactly compute $Z(x)$, we turned to the expectation of $Z(x)$. As shown in [13], the number of configuration pairs at a distance $xN$ is
The solution space structure of random constraint satisfaction problems with growing domains

t(x) = N^{\alpha N} \left( \frac{N}{xN} \right)^{N^\alpha - 1};

and the probability of a pair of configurations being two solutions is written as

\[ q(x) = \{(1-p)^2 + p(1-p)[(1-x)^k + g(x)]\}^{N \ln N}, \]

where

\[ g(x) = \frac{-k(k-1)x(1-x)^{k-1}}{2N}. \]

Then the expectation of \( Z(x) \), denoted as \( \mathbb{E}(Z(x)) \), is the product of \( t(x) \) and \( q(x) \).

Since the domain size grows with \( N \) in model RB, it is convenient to define the normalised version of \( \mathbb{E}(Z(x)) \), given \( k, \alpha, p \) and \( r \), as

\[ f(x) = \lim_{N \to \infty} \frac{\ln(\mathbb{E}(Z(x)))}{(N \ln N)} = \alpha(1+x) + r \ln [(1-p)^2 + p(1-p)(1-x)^k]. \]  (5)

Actually, \( f(x) \) is the annealed entropy density, which decreases as the parameter \( r \) increases. It is easy to see when \( f(x) < 0 \), \( \mathbb{E}(Z(x)) \to 0 \), then by Markov’s inequality

\[ P(Z(x) > 0) \leq \mathbb{E}(Z(x)), \]  (6)

we have w.h.p. \( Z(x) = 0 \).

In figure 1, we plotted \( f(x) \) as a function of \( x \) for \( k = 2, p = 0.4, \alpha = 0.8 \) and several different \( r \) values. The top line has a relatively small \( r \), and we can see that \( f(x) \) is above 0. Note that \( f(x) > 0 \) does not guarantee that there is an exponential number of solution pairs at distance \( xN \), because \( \ln(\mathbb{E}(Z(x))) \) is only an upper bound of the entropy \( \mathbb{E}(\ln(Z(x))) \), given by Jensen’s inequality. As \( r \) increases, the \( f(x) \) curve becomes lower and lower. At a certain value \( \hat{r} = 0.8815 r^* \) in our example in figure 1, \( f(x) \) reaches 0. Beyond \( \hat{r} \), \( f(x) = 0 \) has two solutions \(^8\) until \( r \) reaches \( r^* \). With \( r > r^* \), it has been proved \(^{[13]} \) that w.h.p. there is no solution in the system, which is consistent with what the curve shows: \( f(x) \) becomes negative for any \( x \in [0, 1] \).

Next, we focused on the regime between \( \hat{r} \) and \( r^* \) (the shaded regime in figure 1), where \( f(x) = 0 \) has two solutions, denoted by \( x_1 \) and \( x_2 \). Using the definition of \( f(x) \) in equation (5), and the inequality (6), we can compute the number of solution pairs at the Hamming distance between \( x_1N \) and \( x_2N \) with \( N \to \infty \),

\[ P \left( \sum_{xN = x_1N + 1}^{x_2N - 1} Z(x) > 0 \right) \leq \mathbb{E} \left( \sum_{xN = x_1N + 1}^{x_2N - 1} Z(x) \right) \leq (x_2 - x_1)N \cdot \max_{x \in (x_1, x_2)} \mathbb{E}(Z(x)) \to (x_2 - x_1)N \cdot \max_{x \in (x_1, x_2)} N f(x) N \to 0. \]  (7)

\(^8\) There are at most two solutions, following the concavity of \( f(x) \) shown in the appendix.
The last equation indicates that w.h.p. there are no solution pairs at the Hamming distance between $x_1 N$ and $x_2 N$.

On the other hand, as $N \to \infty$, using the Paley–Zigmund inequality we have

$$P\left[X > \frac{1}{N} \mathbb{E}(X)\right] \geq \left(\frac{\mathbb{E}(X) - \frac{1}{N} \mathbb{E}(X)}{\mathbb{E}(X^2)}\right)^2$$

$$\to \left(1 - \frac{1}{N}\right)^2 \to 1,$$

where we have made use of equation (4), which says $\lim_{N \to \infty} \frac{(\mathbb{E}(X^2))^2}{\mathbb{E}(X^3)} = 1$. The last equation means w.h.p. the number of solutions is bigger than its mean divided by $N$. Then it follows that w.h.p. in the regime where an $x_1$ and $x_2$ pair exists (e.g. the shaded regime in figure 1), the system has an exponential number of solutions and their Hamming distance is discontinuously distributed. In other words, the solution space is clustered. Actually we can show that, for all the parameters of model RB, there always exists such a clustered regime. Proof for the existence of an $x_1$ and $x_2$ pair is given in the appendix.

3.2. Organisation of clusters

In this section, we give a precise description of the clustering of solutions, including the bounds of the diameter of clusters, the distance between clusters, the number of solutions in a given cluster and the number of clusters in the satisfiable phase. Given the result from the previous section, and using the methods from Achlioptas and Ricci–Tersenghi (see [10] section 3 of [28, 29]), we actually have a concrete way to split the solution space and to put solutions into different clusters: assuming that all of the solutions

Figure 1. Annealed entropy density (the leading order of the logarithm of the number of solution pairs $f(x)$, equation (5)), for $\alpha = 0.8$, $p = 0.4$, $k = 2$. From top to bottom, the $r$ values are $0.81 r^*$, $0.8815 r^*$, $0.94 r^*$ and $r^*$, respectively. In the shaded regime $(0.8815 r^*, r^*)$, $f(x) = 0$ has two solutions, denoted by $x_1$ and $x_2$. One example of the two solutions is labelled in the figure for $r = 0.94 r^*$. The last equation indicates that w.h.p. there are no solution pairs at the Hamming distance between $x_1 N$ and $x_2 N$.
are known, we can split the solution space by the curved surface \( \{ y | d(x, y) = x_1 N \} \), and obtain a set of regions \( \mathcal{A} \). In more detail, we can do this as follows:

1. Initialise \( \mathcal{A} = \{ \mathcal{S} \} \), where \( \mathcal{S} \) denotes the set of all solutions.

2. For every solution \( x \in \mathcal{S} \), repeat the splitting step (step 3) around \( x \).

3. Splitting step around \( x \): denote the only region that includes \( x \) in \( \mathcal{A} \) by \( A \). If there is \( y \in A \) that satisfies \( d(x, y) > x_1 N \), then let \( B = \{ y | d(x, y) \leq x_1 N \} \), \( C = A \setminus B \), and let \( \mathcal{A} = (\mathcal{A} \cup \{ B \}) \cup \{ C \} \setminus \{ A \} \).

The final \( \mathcal{A} \) is the set of regions we want. From equation (7) we can show that w.h.p., \( \mathcal{A} \) has the following properties:

- The diameter of each region is at most \( x_1 N \). If there are two solutions at a distance larger than \( x_1 N \) in a region, the splitting step will definitely split them into different regions.

- The distance between each pair of regions is at least \( x_1 N - \frac{1}{x_1 N} \). To show this, we assume there are three solutions \( x, y \), and \( z \), which are put into two clusters after the splitting step around \( x \). \( y \) is put into the same region as \( x \), and \( z \) is put into a different region. Then we have \( d(x, y) \leq x_1 N \), \( d(x, z) \geq x_2 N \), and the triangle rule implies that \( d(y, z) \geq (x_2 - x_1) N \).

Note that \( \mathcal{A} \) is not univocally defined, since it depends on the order of splittings. As the diameter \( x_1 N \) could be bigger than the distance \( (x_2 - x_1) N \), two solutions at the distance between \( (x_2 - x_1) N \) and \( x_1 N \) could be in one or two regions.

Another important property we are interested in is the number of solutions contained in the clusters. For convenience we will omit the term ‘w.h.p.’. From the above analysis, we know that the diameter of each cluster is at most \( x_1 N \), thus the number of solution pairs in a given cluster is bounded above by the total number of solution pairs at distances smaller than \( x_1 N \). Letting

\[
l = \max_{x \in \left[ \frac{1}{N}, x_1 \right]} \mathbb{E}(Z(x))
\]

and using Markov’s inequality, in the large \( N \) limit, we have

\[
P\left( \sum_{x_1 N = 1}^{x_1 N} Z(x) \geq N^2 l \right) \leq \frac{\sum_{x_1 N = 1}^{x_1 N} \mathbb{E}(Z(x))}{N^2 l} \leq \frac{N l}{N^2 l} = \frac{1}{N} \to 0.
\]

We can see that every cluster in \( \mathcal{A} \) has at most \( N^2 l \) pairs of solutions, which implies that every cluster in \( \mathcal{A} \) has at most \( N \sqrt{l} \) solutions. We know that \( f(x) \) is a concave function, and it monotonically decreases as \( x \in [0, x_1] \) (see appendix for proof), thus when \( N \) is large, the upper bound of the number of solutions in one cluster satisfies

\[
N \sqrt{l} < N^{1/2} \alpha + r \ln(1-p) N^{1/2}.
\]

doi:10.1088/1742-5468/2015/12/P12006
Note that, compared with $\frac{1}{N}E(X)$, which is a lower bound of the total number of solutions (equation (8)), the number of solutions in one cluster is exponentially smaller. More precisely, if we define a complexity function $\Sigma$ that represents the leading order of the logarithm of the number of clusters divided by $N\log N$ (in our system, the correct scaling for densities is $N\log N$), we have

$$\Sigma \geq \frac{1}{N\ln N} \ln \left( \frac{\frac{1}{N}E(X)}{N\sqrt{l}} \right)$$

$$\geq \frac{1}{2} \left[ \alpha + r\ln(1 - p) - \frac{2}{N} \right].$$

The last equation implies that in the satisfiable phase, the complexity is positive all the way down to the satisfiability transition. A direct implication of the above results is that in the whole parameter range, the phase diagram of model RB does not contain a condensed clustered phase, because one cannot find a finite number of clusters that contain almost all of the solutions. In replica symmetry breaking theory, the existence of a clustering phase is indicated by $\tilde{\Sigma}(m = 1) > 0$ and the existence of the condensation phase is indicated by $\tilde{\Sigma}(m = 1) < 0$, where $\tilde{\Sigma}$ denotes the complexity, which is the leading order of the logarithm of the number of pure states as a function of the Parisi parameter $m$ [6]. If the Parisi parameter $m = 1$, the first step replica symmetry breaking solution gives equal weights to each pure state, thus the total free energy is identical to the replica symmetry free energy. We can see that our definition of complexity $\Sigma$ is very similar to $\tilde{\Sigma}(m = 1)$ since it also gives equal weights to different clusters. Thus, $\Sigma > 0$ all the way down to the satisfiability transition is another way to show that there is no condensation transition in model RB. Note that, since our definition of clusters is different to that of pure states (as we do not refer to properties of the Gibbs measure), our claim is not a proof.

4. Conclusion and discussion

In this paper, we described in detail the solution space structure of the model RB problem using rigorous methods. We showed that close to the satisfiability transition, solutions are clustered into an exponential number of clusters, each of which contains a sub-exponential number of solutions. We showed that there is no condensation transition in model RB, which supports the conclusions of Zhao et al [19], who used non-rigorous cavity methods derived from statistical physics.

We note that the factor graph of model RB has a special property that the average degree of variables is very large (which increases with the number of variables $N$), which is the same as the $K$-SAT problem with increasing $K$ [30]. We believe that this property might be another reason to explain the absence of condensation. We will perform future investigations to develop a better understanding of this point.

In this article, we proved that solutions cluster close to the satisfiability transition; however, our results do not precisely determine where the clustered phase...
begins. Heuristically, the clustering transition can be estimated when the one-step-replica-symmetry-breaking cavity method (with a Parisi parameter of $m = 1$) begins to have a non-trivial solution. We will address this point in a future work.

It has been shown that instead of clustering, the freezing of clusters is the real reason for algorithmic hardness. Numerical experiments performed by [19] and [18] showed that when starting from $\hat{r}$ (where $f(x) = 0$ has only one solution), even the most efficient algorithms begin to fail at finding solutions, thus suggesting that clusters become immediately frozen at $\hat{r}$. This would be interesting to study in detail.

Acknowledgments

PZ wishes to thank Christopher Moore for helpful conversations.

Appendix. Concavity of $f(x)$ and existence of solution pairs

The first and second derivatives of $f(x)$ with respect to $x$ read as

$$\frac{\partial f}{\partial x} = \alpha - \frac{rpk(1-x)^{k-1}}{1-p + p(1-x)^k},$$

$$\frac{\partial^2 f}{\partial x^2} = \frac{rpk}{[1-p + p(1-x)^2]} [(k-1)(1-p)(1-x)^{k-2} - p(1-x)^{2k-2}].$$

Then it is easy to check that $\frac{\partial^2 f}{\partial x^2}$ is always positive for $x \in [0, 1]$ with $k \geq 2$ and $p < 1 - \frac{1}{k}$, which implies the concavity of $f(x)$.

Observe that both $f(0)$ and $f(1)$ are positive in the satisfiable phase, $f(0) = f(1) = 0$ at the satisfiability transition, and $\frac{\partial f(0)}{\partial x} = \alpha > 0$. Thus, using the concavity of $f(x)$, it is obvious that there must exist $r < r^*$, and a pair of $x_1, x_2$, such that $f(x) < 0$ with $x_1 < x < x_2$. Moreover in the satisfiable phase, since $f(0) = \alpha + r\ln(1-p) > 0$, and $x_1$ is the first point where $f(x)$ reaches 0, we can conclude that $f(x)$ is a monotonically decreasing function as $x \in [0, x_1]$.

References

[1] Kirkpatrick S and Selman B 1994 Science 264 1297–301
[2] Monasson R, Zecchina R, Kirkpatrick S, Selman B and Troyansky L 1999 Nature 400 133–7
[3] Mézard M, Parisi G and Zecchina R 2002 Science 297 812
[4] Mézard M and Zecchina R 2002 Phys. Rev. E 66 056126
[5] Mézard M and Parisi G 2001 Eur. Phys. J. B 20 217–33
[6] Krzakala F, Montanari A, Ricci-Tersenghi F, Semerjian G and Zdeborova L 2007 Proc. Natl Acad. Sci. USA 104 10318
[7] Montanari A, Ricci-Tersenghi F and Semerjian G 2008 J. Stat. Mech. P04004
[8] Zdeborova L and Krzakala F 2007 Phys. Rev. E 76 031131
[9] Mézard M, Mora T and Zecchina R 2005 Phys. Rev. Lett. 94 197205
[10] Achlioptas D, Coja-Oghlan A and Ricci-Tersenghi F 2011 Random Struct. Algorithms 38 251
[11] Achlioptas D and Coja-Oghlan A 2008 Proc. 49th FOCS pp 793–802
[12] Coja-Oghlan A and Zdeborova L 2012 Proc. 23rd SODA pp 241–50
[13] Xu K and Li W 2000 J. Artif. Intell. Res. 12 93–103

doi:10.1088/1742-5468/2015/12/P12006
The solution space structure of random constraint satisfaction problems with growing domains

[14] Smith B and Dyer M 1996 Artif. Intell. 81 155
[15] Xu K and Li W 2006 Theor. Comput. Sci. 355 291
[16] Xu K, Boussemart F, Hemery F and Lecoutre C 2007 Artif. Intell. 171 514
[17] Liu T, Lin X and Wang C 2011 Proc. 22nd IJCAI pp 611–6
[18] Zhao C Y, Zhou H J, Zheng Z M and Xu K 2011 J. Stat. Mech. P02019
[19] Zhao C Y, Zhang P, Zheng Z M and Xu K 2012 Phys. Rev. E 85 016106
[20] Achlioptas D and Moore C 2006 SIAM J. Comput. 36 740–62
[21] Lecoutre C 2009 Constraint Networks: Techniques and Algorithms (London: ISTE and Hoboken: Wiley)
[22] Smith B 2001 Theor. Comput. Sci. 265 265–83
[23] Frieze A and Molloy M 2006 Random Struct. Algorithms 28 323–39
[24] Fan Y and Shen J 2011 Artif. Intell. 175 914–27
[25] Fan Y, Shen J and Xu K 2012 Artif. Intell. 193 1–17
[26] Zhang P, Zeng Y and Zhou H 2009 Phys. Rev. E 80 021122
[27] Barbier J, Krzakala F, Zdeborova L and Zhang P 2013 J. Phys.: Conf. Ser. 473 012021
[28] Achlioptas D and Ricci-Tersenghi F 2006 Proc. STOC’06 pp 130–9
[29] Achlioptas D 2008 Eur. Phys. J. B 64 395
[30] Frieze A M and Wormald N C 2005 Combinatorica 25 297–305

doi:10.1088/1742-5468/2015/12/P12006