Physica A 537 (2020) 122708

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Clustering phase of a general constraint satisfaction problem model *d*-*k*-CSP

Wei Xu [a], Fuzhou Gong [b], Guangyan Zhou [c,*]

[a] School of Mathematics and Physics, University of Science and Technology Beijing, Beijing 100083, China
[b] Institute of Applied Mathematics, Academy of Mathematics and System Science, Chinese Academy of Sciences, Beijing 100190, China
[c] Department of Mathematics, Beijing Technology and Business University, Beijing 100048, China

ARTICLE INFO

Article history:
Received 8 April 2019
Received in revised form 6 August 2019
Available online 18 September 2019

Keywords:
Constraint satisfaction problem
Solution space structure
Clustering phase transition
Problem hardness
Belief propagation

ABSTRACT

Relation between problem hardness and solution space structure is an important research aspect. Model *d*-*k*-CSP is a random model of constraint satisfaction problem, which generates very hard instances when $r = 1$ or r is near 1, where r represents normalized constraint density. We study the clustering phase of the model, and find that if r is below and close to 1, the solution space contains many widely distributed and well-separated small frozen clusters, which should be the reason why the generated instances are hard to solve.

1. Introduction

The constraint satisfaction problem, or CSP, is an important topic in the interdisciplinary area of statistical physics, computer science and information theory. In the theoretical aspect, CSPs play a significant role in understanding the problem hardness as well as properties of disordered systems. In practical use, CSPs are applied to many tasks, such as timetabling, hardware configuration, transportation scheduling, factory scheduling, floorplanning, error-correcting code, etc. A CSP formula contains variables and constraints, where each variable can be assigned a value from its domain, and the constraints contain a set of variables called the constraint scope that restricts their allowed joint values. One elementary question is to assign all variables such that all constraints are satisfied simultaneously; meanwhile, another elementary question is to determine whether a solution exists.

Popular CSPs, such as *k*-SAT and Coloring, have been intensely studied, and fruitful results have been obtained. Cheeseman et al. in [1], found that many CSPs undergo a satisfiable–unsatisfiable transition: when the constraint density is low, almost every instance has a solution; but when the constraint density increases and exceeds an inherent point, the solutions suddenly disappear for almost every instance. Friedgut et al. supported this opinion and proved that satisfiability of K-SAT goes through a sharp threshold (a concept from the random graph category, which is weaker than the phase transition expression) [2]. Then the lower and upper bounds of the satisfiable–unsatisfiable transition point were gradually improved in several papers. Mézard et al. applied the cavity method which was derived from spin-glass research to the CSP and separated the satisfiable region into two phases, the replica symmetric (RS) phase, where all the solutions belong to a single state (cluster), and the one step replica symmetry breaking (1RSB) phase, where the solutions belong to many states [3–6]. The transition from the RS phase to the 1RSB phase is called the dynamical transition (or clustering

[*] Corresponding author.
E-mail address: zhouguangyan@btbu.edu.cn (G. Zhou).

https://doi.org/10.1016/j.physa.2019.122708
0378-4371/© 2019 Elsevier B.V. All rights reserved.

transition). Further studies using the cavity method found the condensation transition, where the solution space starts to be dominated by a few large clusters [7,8]. Condensation transition separated the 1RSB into two phases, dynamical 1RSB phase (d1RSB) and static 1RSB phase (s1RSB). Then freezing transition was found, where a linear number of frozen variables (fixed throughout the cluster) arise in almost every cluster [9]. It is worth mentioning that inspired by the 1RSB type cavity method, the survey inspired decimation (SID) algorithm [4,5] was proposed and is known to be a highly efficient algorithm when the constraint density is rather close to the satisfiable–unsatisfiable transition.

Studies show that solution space structures affect problem hardness and the performance characteristics of different algorithms. Problems attain their maximum hardness when the constraint density approaches the satisfiable–unsatisfiable transition. Dynamical transition may have impact on energy relaxation strategy such as Monte Carlo simulated annealing algorithm. Experiments suggest that belief propagation guided decimation (BPD) algorithm is effective up to the condensation transition [7]. It was stated in [10] that the intuitive assumption supporting the SID validity is that many clusters exist in the solution space, whereas for BPD, it is assumed that most solutions belong to one cluster. Many research works suggest that the freezing transition is a pivotal transition for algorithm validity. To find out the relation between problem hardness and solution space structure, many outstanding algorithms have been designed, such as Monte Carlo algorithms [11], stochastic local search algorithms FMS [12] and ASAT [13], message passing algorithm BSP [14], etc. All those algorithms can find solutions efficiently beyond the dynamical threshold, and close to the satisfiable–unsatisfiable transition.

Random CSP model, also known as random CSP instances generator, is proposed to enrich the study of the CSPs. The initial proposed random models [15,16] are called models A, B, C, and D, where the constraint scope size and domain size are fixed. Unfortunately for these models, it was proved by Achlioptas et al. [17] that the generated instances suffer from trivial unsatisfiability, that is, almost all instances are unsatisfiable when the number of variables is large. To overcome this flaw, many alternative models have been proposed. One technique is to incorporate a special combinatorial structure on constraints and ensure that the generated instance has certain consistency properties [15,18]. Another technique is to change the scales of the parameters, including the size of the domain and the length of the constraint scope [19–23]. In this work, we study the model d-k-CSP which proposed in [23], where the constraint scope length and/or the domain size are increasing with the number of variables.

Model RB, a special case of model d-k-CSP, also plays a significant role in computer science. It can be used to generate very hard benchmark instances [24]. In [25] and [26], solution space structure of model RB was studied, it was proved that the clustering phase exists and persists until the satisfiable–unsatisfiable transition point, so the condensation phase does not exist. Their method was based on the study of the number of solution pairs at different distances. Mézard et al. [27] studied the clustering phenomenon by counting the number of solution pairs, then Achlioptas et al. found a way to prove the existence of clustering phase for K-SAT [28–31].

Some other models do not have condensation phase (or s1RSB phase), such as k-XORSAT and the locked model [32]. If s1RSB does not appear, after the dynamical transition, clusters become frozen until no clusters exist. In the d1RSB phase, there are many clusters, each of which has a Bethe measure, but the measure of all solutions is also a Bethe measure. Belief propagation (BP) still converges and reveals the distribution of solutions in the d1RSB phase. We perform BP for the d-k-CSP model, and find that it always converges in the region $r < 1$, so there should exists either RS phase or d1RSB phase. The locked model has a special solution space structure, that the solutions are all isolated from each other, which is different from the d-k-CSP model. The d1RSB phase of k-XORSAT should be similar with the d1RSB phase of d-k-CSP.

In this paper, using the same method as in [26], we prove that the clustering phase exists before the satisfiable–unsatisfiable transition. There are many well-separated cluster-regions (a cluster-region is a union of some clusters), and the diameter of a cluster decreases to be very small when r approaches 1. Marginals obtained from Bethe–Peierls approximation show that the clusters distribute widely in the solution space. Whitening procedure on solutions shows that variables become frozen before $r = 1$. It concludes that when r is below and close to 1, the solution space contains many widely distributed well-separated small frozen clusters. When r is below and close to 1, the problem is hard to solve, so the above structure should be the reason of the hardness. As d-k-CSP is an important model of the CSP, our result will provide a further understanding of the relation between solution space structure and complexity.

The rest of the paper is organized as follows: we give a definition of the model d-k-CSP in Section 2, and describe the method of showing clusters in Section 3, then we prove the clustering phenomenon in Sections 4, 5, 6, and a special case will be discussed in Section 7, marginals of variables and distribution of clusters are studied by physical methods in Section 8, then whitening procedure and frozen variables are discussed in Section 9.

2. Model d-k-CSP and definitions

Fan, Shen and Xu (2012) proposed a general model of a random CSP with varying constraint scope length and varying domain size, called model d-k-CSP [23]. An instance of model d-k-CSP is composed of a set of variables $V = \{x_1, x_2, \ldots, x_n\}$ and a set of constraints $C = \{C_1, C_2, \ldots, C_t\}$, where n, t are the number of variables and constraints respectively. Each variable $x_i (i = 1, \ldots, n)$ can only be assigned a value from its domain D_i, where every cardinality $|D_i| = d$, and $d \geq 2$ is a function of n. Each constraint $C_i (i = 1, \ldots, t)$ is a pair (X_i, R_i), where $X_i = (x_{i_1}, x_{i_2}, \ldots, x_{i_k})$ is the constraint scope, $k = k(n) \geq 2$ is the constraint scope length, and $R_i \subseteq D_{i_1} \times D_{i_2} \times \cdots \times D_{i_k}$ is a set of compatible tuples of values. A d-k-CSP instance I is generated by the following two steps:

Table 1
Other conditions of different cases.

The cases	Other conditions
Case A	$d(n) = const, k(n) \to \infty, k(n) = o(n)$
Case B	$d(n) = const, k(n) \to \infty, k(n) = bn, 0 < b < 1$
Case C	$d(n) \to \infty, k(n) \to \infty, k(n) = o(n)$
Case D	$d(n) \to \infty, k(n) \to \infty, k(n) = bn, 0 < b < 1$
Model RB [26]	$k(n) = const, d(n) = n^{\alpha}$

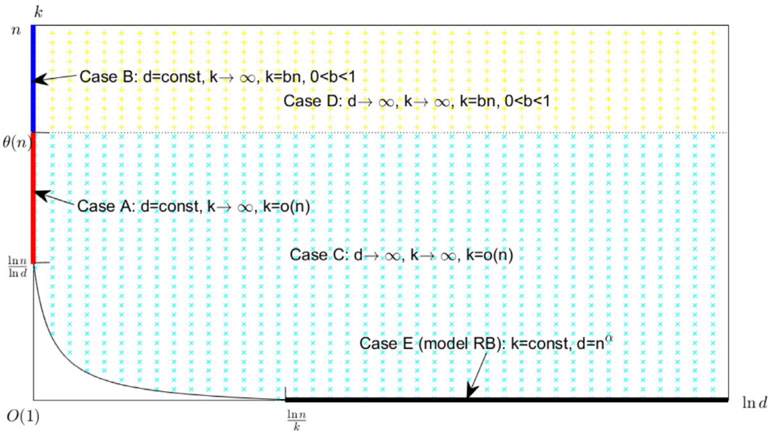
k
n
Case B: d=const, k→ ∞, k=bn, 0<b<1
Case D: d→ ∞, k→ ∞, k=bn, 0<b<1
θ(n)
Case A: d=const, k→ ∞, k=o(n)
Case C: d→ ∞, k→ ∞, k=o(n)
$\frac{\ln n}{\ln d}$
Case E (model RB): k=const, d=n^α
O(1)
$\frac{\ln n}{k}$
ln d

Fig. 1. The cases in the coordinate system of $\ln d$ and k. On the k axis, there are Case A and Case B. On the $\ln d$ axis, there is the case of model RB. In the shadowed area there are Case C and Case D.

1. Select t constraints randomly with repetition. Each constraint is formed by selecting k of the n variables randomly without repetition.
2. For each constraint, select $q = (1-p)d^k$ compatible tuples of values randomly without repetition, where $0 < p < 1$ is a constant.

A constraint $C_i = (X_i, R_i)$ is said to be satisfied by an assignment $\sigma \in D^n$ if the values assigned to X_i are in the set R_i. An assignment σ is a solution if it satisfies all constraints. A CSP instance is called satisfiable if there are solutions, and called unsatisfiable otherwise. Fan et al. [23] proved that model d-k-CSP has a satisfiable–unsatisfiable transition: assume that $r > 0$ and $0 < p < 1$ are constants, $t = r\frac{n \ln d}{-\ln(1-p)}$, $\lim_{b\to\infty} \frac{1}{d}$ exists, $k \geq \frac{1}{1-p}$ and there is a positive real number ϵ such that $k \ln d \geq (1+\epsilon)\ln n$; then,

$$\lim_{n\to\infty} Pr[\text{I is satisfiable}] = \begin{cases} 1 & r < 1, \\ 0 & r > 1. \end{cases}$$

As part of the proof, they found that in the satisfiable phase $r < 1$,

$$\lim_{n\to\infty} \frac{(\mathbb{E}(X))^2}{\mathbb{E}(X^2)} = 1. \tag{1}$$

We sort the model into different cases based on the speeds of which the domain size $k(n)$ and/or the length of constraint scope $d(n)$ grow with the number of variables n. First of all, in order to guarantee the existence of the satisfiable–unsatisfiable transition, the conditions in Theorem 2.1 of [23] should be satisfied. Then we consider other conditions in Cases A, B, C, D and model RB in Table 1. Taking Case A for an example, the parameter $d(n)$ is a constant, $k(n)$ tends to infinity, and $k(n)$ is an infinitesimal of higher order than n. In Fig. 1, we show the conditions of the cases in the coordinate system of $\ln d$ and k.

A solution pair is a sequence of two assignment solutions. The (Hamming) distance between two solutions is the number of variables being assigned different values by the two solutions. For a random d-k-CSP instance, the number of solution pairs at distance $x(nx = 1, 2, \dots, n)$ is denoted by $Z(x)$, and $\mathbb{E}(Z(x))$ is its expectation in the model. *Cluster* is the connected component in the solution space, where every pair of solutions are considered to be adjacent if they are at distance 1. *Cluster-region* is a union of some clusters. The *distance between two cluster-regions* is the minimum distance of their solution pairs not belonging to the same cluster-region. The *diameter* of a cluster-region is the maximum distance of two solutions in the cluster-region. The *clustering phase* describes the phase where the solution space breaks apart into an exponential number of well-separated clusters, with each cluster containing a sub-exponential number of solutions. The *condensation phase* describes the phase where a finite number of clusters contain almost all of the solutions, which is different from the clustering phase.

3. Method to show the existence of clusters

The method used in this work is based on the distances among the solutions [26–31]. If solution-pairs at distances between αn and βn do not exist, the solution space can be split into cluster-regions, with the cluster-region diameter smaller than αn and the distance among cluster-regions larger than $\beta n - \alpha n$. With this method, to obtain the clustering properties there are two steps left.

Firstly, we should determine the values of α and β such that w.h.p. solution-pairs at distances between αn and βn do not exist, where "w.h.p." means that the probability of an event tends to 1 as $n \to \infty$. For this purpose, we only need to prove that there exists a positive δ such that $\sup_{n\to\infty} f_0(x) < -\delta$ for $\alpha < x < \beta$, where the $f_0(x)$ is defined by $\ln \mathbb{E}(Z(x))/n$. If this is satisfied, by the first moment method, we have

$$\sup_{n\to\infty} P\left(\sum_{xn=\alpha n}^{\beta n} Z(x) > 0\right) \le \sup_{n\to\infty} \mathbb{E}\left(\sum_{xn=\alpha n}^{\beta n} Z(x)\right) = \sup_{n\to\infty} \sum_{xn=\alpha n}^{\beta n} e^{nf_0(x)} \le \sup_{n\to\infty} ne^{-\delta n} = 0,$$

which means w.h.p. solution-pairs at distances between αn and βn do not exist. Or we only need to prove that $\sup_{n\to\infty} g_0(x) < -\delta$ for $\alpha < x < \beta$, where $g_0(x)$ is defined by $\ln \mathbb{E}(Z(x))/(n \ln d)$.

Secondly, we should estimate the number of cluster-regions and the number of solutions in each cluster-region. Let $l = \max_{0<xn\le\alpha n} \mathbb{E}(Z(x))$, by the first moment method and the definition of l, we obtain

$$P\left(\sum_{xn=1}^{an} Z(x) \ge n^2 l\right) \le \frac{\sum_{xn=1}^{an} \mathbb{E}(Z(x))}{n^2 l} \le \frac{nl}{n^2 l} \le \frac{1}{n} \to 0,$$

that is, w.h.p., the number of solution-pairs in each cluster-region is smaller than $n^2 l$. Thus, the number of solutions in each cluster-region is smaller than $nl^{0.5}$. We can give a lower bound of the number of all solutions; by using the Paley–Zigmund inequality and Eq. (1), we have

$$P[X > \frac{1}{n}\mathbb{E}(X)] \ge \frac{\left(\mathbb{E}(X) - \frac{1}{n}\mathbb{E}(X)\right)^2}{\mathbb{E}(X^2)} \to (1 - \frac{1}{n})^2 \to 1.$$

The number of cluster-regions must be larger than the lower bound of the number of all solutions divided by the upper bound of the number of solutions in each cluster-region. This is to say that the number of cluster-regions is larger than

$$\frac{\frac{1}{n}\mathbb{E}(X)}{nl^{0.5}} = \frac{\frac{1}{n}d^n(1-p)^t}{nl^{0.5}}. \tag{2}$$

To give a lower bound of Eq. (2) in the limit condition of $n \to \infty$, we only need to give an upper bound of $\sup_{n\to\infty} l$.

To summarize,

- Given $\delta > 0$, we should find α and β that for $\alpha < x < \beta$, $\sup_{n\to\infty} f_0(x) < -\delta$ or $\sup_{n\to\infty} g_0(x) < -\delta$.
- We should give an upper bound to $\sup_{n\to\infty} l$ to estimate the number of cluster-regions.

4. Find the number of solution-pairs and the values of α, β

The expression of $\mathbb{E}(X)$ should be given first. Using the same analysis as in Eqs. (8) and (9) in article [21], we have

$$\mathbb{E}(Z(x)) = d^n C_n^{nx}(d-1)^{nx}\left[\frac{C_{d^k-1}^q}{C_{d^k}^q}\sigma(x) + \frac{C_{d^k-2}^q}{C_{d^k}^q}(1-\sigma(x))\right]^t, \tag{3}$$

where $\sigma(x) = C_{n-nx}^k / C_n^k$, the C_i^j is combination formula $C_i^j = \frac{i!}{j!(i-j)!}$, and C_i^j takes the value of 0 by definition if $i < j$. Parameters n, d, k, q, and t are all from the definition of model d-k-CSP, and d, k, and x are actually functions of n, that is, $d = d(n)$, $k = k(n)$, $x = x(n)$. We consider the limit situation of $n \to \infty$. By simplification and asymptotic estimation, and since $d^k \to 0$ in all the cases we study here, thus

$$\frac{C_{d^k-1}^q}{C_{d^k}^q} = 1 - p,\ \frac{C_{d^k-2}^q}{C_{d^k}^q} = (1-p)^2 + O(d^{-k}) \to (1-p)^2.$$

In the following of this section, we focus on the signs of $f_0(x) = \ln(\mathbb{E}(Z(x)))/n$ and $g_0(x) = \ln(\mathbb{E}(Z(x)))/(n \ln d)$. When $d = const$, $f_0(x)$ tends to a finite value, and

$$f_0(x) \to A(x) + B(x),$$

where

$$A(x) = 1/n(\ln C_n^{nx} + nx\ln(d-1)),$$

$$B(x) = \ln d - r \ln d + r \frac{\ln d}{-\ln(1-p)} \ln\left(1 - p + p\sigma(x)\right).$$

Also note that when $0 < x < 1$ is a constant, by the Stirling formula, as $n \to \infty$,

$$\ln C_n^{nx} \to -n \ln(x^x (1-x)^{1-x}).$$

When $d \to \infty$, we study $g_0(x)$ instead because it tends to a finite value, and

$$g_0(x) \to C(x) + D(x),$$

where

$$C(x) = x,$$

$$D(x) = 1 - r + r \frac{1}{-\ln(1-p)} \ln\left(1 - p + p\sigma(x)\right).$$

Note that parameter $d \geq 2$ and $0 \leq \sigma(x) \leq 1$, so we have $A(x)$ is positive and $\ln(1-p) \leq \ln\left(1 - p + p\sigma(x)\right) \leq 0$. Therefore if $r < 0.5$, we have

$$A(x) + B(x) > \ln d - 2r \ln d > 0,$$

$$\lim_{n\to\infty} f_0(x) \geq 0,$$

and also

$$\lim_{n\to\infty} g_0(x) \geq 0.$$

This is to say, that if $r < 0.5$, we cannot find α, β such that for $\alpha < x < \beta$ $\lim_{n\to\infty} f_0(x) < 0$ or $\lim_{n\to\infty} g_0(x) < 0$.

But when $0.5 < r < 1$, we can find the α, β for Cases A, B, C, and D. For each case we give three statements in the following:

(1) For $0 < x < \alpha$, we have $\inf_{n\to\infty} f_0(x) \geq 0$ or $\inf_{n\to\infty} g_0(x) \geq 0$.

(2) For $\beta < x < \beta^*$ (where β^* is some constant bigger than β), we have $\inf_{n\to\infty} f_0(x) \geq 0$ or $\inf_{n\to\infty} g_0(x) \geq 0$.

(3) For $\alpha(1+\epsilon) < x < \beta(1-\epsilon)$ (where ϵ is arbitrarily small), we have $\sup_{n\to\infty} f_0(x) < -\delta$ or $\sup_{n\to\infty} g_0(x) < -\delta$ (δ is positive and depends on ϵ).

4.1. Case a where $d = const$, $k \to \infty$, and $k = o(n)$

(1) If $x \in (0, a_0/k)$, where $a_0 > 0$ is a constant defined in the following, then we have $\inf_{n\to\infty} f_0(x) \geq 0$.

Since in this case $k \to \infty$, then for any constant a and $xn \in (0, an/k)$, we have $x \to 0$ and $f_0(x) \to B(x)$. If $x = a/k$, where $a = const$, we have

$$\sigma(x) = C_{n-nx}^k / C_n^k = \frac{(n-nx)...(n-nx-k+1)}{n...(n-k+1)} < (1 - \frac{nx}{n})^k \to e^{-xk} = e^{-a},$$

$$\sigma(x) = C_{n-nx}^k / C_n^k > (1 - \frac{nx}{n-k+1})^{\frac{n-k+1}{nx} \frac{nx}{n-k+1} k} \to e^{-\frac{nx}{n-k+1} k} \to e^{-a}.$$

Combining the above two inequalities, we obtain $\sigma(x) \to e^{-a}$. Then we find $B(a/k) \to f_1(a)$, where $f_1(a)$ is defined by

$$f_1(a) \triangleq \ln d - r \ln d + r \frac{\ln d}{-\ln(1-p)} \ln\left(1 - p + pe^{-a}\right),$$

which is decreasing with a. Solving equation $f_1(a) = 0$ with respect to variable a, we obtain the solution

$$-\ln \frac{(1-p)^{1/r} - (1-p)^2}{1 - p - (1-p)^2} \triangleq a_0,$$

which is a positive constant when $0.5 < r < 1$. Function $f_0(x) \to B(x)$, $B(x)$ is a decreasing function, and $B(a_0/k) \to f_1(a_0) = 0$; therefore, we find that when $x < a_0/k$, $\inf_{n\to\infty} f_0(x) \geq 0$.

(2) If $x \in (b_0, b_1)$, where b_0 and b_1 are constants defined in the following, we have $\inf_{n\to\infty} f_0(x) \geq 0$.

If x is a positive constant, then $\sigma(x) = C_{n-nx}^k / C_n^k \to 0$, and

$$f_0(x) \to \ln d - \ln(x^x (1-x)^{1-x}) + x \ln(d-1) - 2r \ln d \triangleq f_2(x). \tag{4}$$

We can draw a picture of function $f_2(x)$, as shown in Fig. 2. The first and second-order derivatives of $f_2(x)$ are

$$f_2'(x) = -\ln x + \ln(1-x) + \ln(d-1),$$

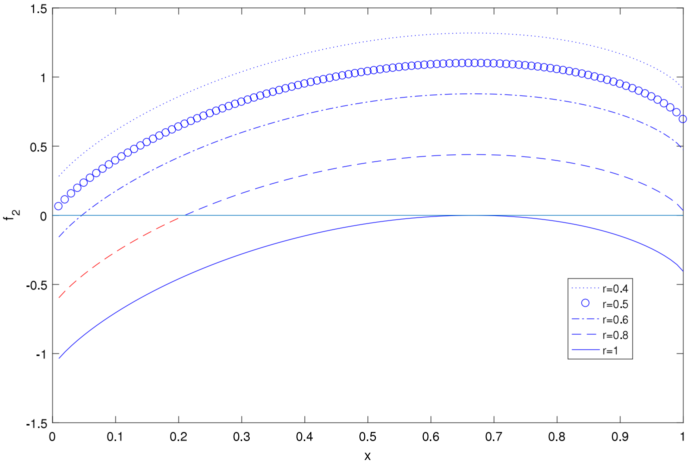

Fig. 2. Function f_2 for $d = 3$ and $r = 0.4, 0.5, 0.6, 0.8, 1$ from top to bottom. There is a region satisfying $f_2(x) < 0$ when $r > 0.5$. Taking $r = 0.8$ as an example, in the region $x \in (0, 0.2)$, $f_2(x) < 0$ (plotted in red).

$$f_2''(x) = -\frac{1}{x} - \frac{1}{1-x} < 0.$$

Clearly, $f_2(x)$ is a concave function. Let $f_2'(x) = 0$, we then obtain $x = \frac{d-1}{d}$. Thus, $f_2(x)$ achieves its maximum value at

$$x = \frac{d-1}{d} \triangleq b_1,$$

and the maximum value is

$$f_2(b_1) = 2(1-r)\ln d.$$

When $0.5 < r < 1$,

$$f_2(b_1) = 2(1-r)\ln d > 0,$$

$$f_2(x=0) = (1-2r)\ln d < 0,$$

so equation $f_2(x) = 0$ has one solution in region $(0, \frac{d-1}{d})$, denoted by b_0. Additionally, there is a region (b_0, b_1) where $f_2(x) > 0$, and since $f_0(x) \to f_2(x)$, we obtain $\inf_{n\to\infty} f_0(x) > 0$.

(3) For arbitrarily small positive number ϵ, when $x \in ((a_0+\epsilon)/k, b_0-\epsilon)$, there exists $\delta > 0$ s.t. $\sup_{n\to\infty} f_0(x) < -\delta$.

Note that $A(x)$ is very small when x is a very small constant, so for $-\frac{1}{2}f_1(a_0+\epsilon)$, which is a positive number, there exists a constant x_0 such that $x_0 < b_0 - \epsilon$, $x_0 < 1/2$ and

$$A(x_0) < -\frac{1}{2}f_1(a_0+\epsilon).$$

Based on this, we let

$$((a_0+\epsilon)/k, b_0-\epsilon) = ((a_0+\epsilon)/k, x_0] \cup (x_0, b_0-\epsilon).$$

In the first range $((a_0+\epsilon)/k, x_0]$, $A(x)$ increases and therefore $A(x) < -\frac{1}{2}f_1(a_0+\epsilon)$; $B(x)$ decreases, thus $B(x) < B((a_0+\epsilon)/k) \to f_1(a_0+\epsilon)$. Also, we have

$$\sup_{n\to\infty} f_0(x) \le -\frac{1}{2}f_1(a_0+\epsilon) + f_1(a_0+\epsilon) = \frac{1}{2}f_1(a_0+\epsilon).$$

In the second range $(x_0, b_0-\epsilon)$, we have $f_0(x) \to f_2(x) < f_2(b_0-\epsilon)$, and therefore

$$\sup_{n\to\infty} f_0(x) \le f_2(b_0-\epsilon).$$

Let $-\delta = \max\{\frac{1}{2}f_1(a_0+\epsilon), f_2(b_0-\epsilon)\}$, then for $x \in ((a_0+\epsilon)/k, b_0-\epsilon)$ we have $\sup_{n\to\infty} f_0 < -\delta$.

4.2. Case B where $d = const$, $k \to \infty$, $k = bn$, and $0 < b < 1$

(1) If $x \in (0, a_1/n)$, where $a_1 > 0$ is a constant defined in the following, we have $\inf_{n\to\infty} f_0(x) \ge 0$.

If $x = a/n$, where a is a positive constant integer, we have

$$\sigma(x = a/n) = \frac{C_{n-a}^{bn}}{C_n^{bn}} = \frac{(n-a)...(n-a-bn+1)}{n...(n-bn+1)} = \frac{(n-bn)...(n-bn-a+1)}{n...(n-a+1)} \rightarrow (1-b)^a;$$

then we find $f_0(a/n) \rightarrow f_3(a)$, where $f_3(a)$ is defined by

$$f_3(a) \triangleq \ln d - r\ln d + r\frac{\ln d}{-\ln(1-p)} \ln\left(1-p+p(1-b)^a\right).$$

Solving $f_3(a) = 0$ with respect to variable a, we obtain the solution

$$\frac{\ln((1-p)^{\frac{1-r}{r}} - 1 + p) - \ln p}{\ln(1-b)} \triangleq a_1.$$

When $0.5 < r < 1$, we have $0 < \ln((1-p)^{\frac{1-r}{r}} - 1 + p) < p$, and thus a_1 is a positive constant. When $x < a_1/n, f_0(x) \rightarrow B(x)$, $B(x)$ is a decreasing function and $B(a_1/n) \rightarrow f_3(a_1) = 0$, so $\inf_{n\rightarrow\infty} f_0(x) \geq 0$.

(2) As the same with Case A, if x is a positive constant, $f_0(x) \rightarrow f_2(x)$. Then for b_0, b_1 and $x \in (b_0, b_1)$, we have $\inf_{n\rightarrow\infty} f_0(x) \geq 0$.

(3) If $x \in ((a_1+\epsilon)/n, b_0-\epsilon)$ where ϵ is arbitrarily small positive number, there exists $\delta > 0$ s.t. $\sup_{n\rightarrow\infty} f_0(x) < -\delta$.

Note that $-\frac{1}{2}f_3(a_1+\epsilon)$ is a positive number, then it is the same with case A, and there exists a constant x_1 such that $x_1 < b_0 - \epsilon$, $x_1 < 1/2$ and

$$A(x_1) < -\frac{1}{2}f_3(a_1+\epsilon).$$

Based on this x_1, we divide the range $((a_1+\epsilon)/n, b_0-\epsilon)$ into

$$((a_1+\epsilon)/n, x_1] \cup (x_1, b_0-\epsilon).$$

In the range $((a_1+\epsilon)/n, x_1]$, $A(x)$ increases and therefore $A(x) < -\frac{1}{2}f_3(a_1+\epsilon)$; $B(x)$ decreases, so $B(x) < B((a_1+\epsilon)/n) \rightarrow f_3(a_1+\epsilon)$. We have in this range

$$\sup_{n\rightarrow\infty} f_0(x) \leq \frac{1}{2}f_3(a_1+\epsilon).$$

In the range $(x_1, b_0-\epsilon)$, we have $f_0(x) \rightarrow f_2(x) < f_2(b_0-\epsilon)$, and thus $\sup_{n\rightarrow\infty} f_0(x) \leq f_2(b_0-\epsilon)$.

Let $-\delta = \max\{\frac{1}{2}f_3(a_1+\epsilon), f_2(b_0-\epsilon)\}$; we have for $x \in ((a_1+\epsilon)/n, b_0-\epsilon)$, $\sup_{n\rightarrow\infty} f_0 < -\delta$.

4.3. Case C where $d \rightarrow \infty$, $k \rightarrow \infty$, and $k = o(n)$

(1) If $x = a/k$, where $a = const$, we have $\sigma(x) \rightarrow e^{-a}$,

$$g_0(x) \rightarrow 1 - r + r\frac{1}{-\ln(1-p)} \ln\left(1-p+pe^{-a}\right) \triangleq g_1(a).$$

Solving equation $g_1(a) = 0$ with respect to variable a gives the solution a_0. Similar to case A, when $x \in (0, a_0/k)$, we have $\inf_{n\rightarrow\infty} g_0(x) \geq 0$.

(2) If x is a positive constant, then $\sigma(x) \rightarrow 0$, and

$$g_0(x) \rightarrow 1 + x - 2r \triangleq g_2(x). \tag{5}$$

We can draw an illustration of function g_2 with respect to variable x in Fig. 3. Equation $g_2(x) = 0$ has one solution, which is $2r - 1$. $g_2(x)$ increases, so we let $b_2 = 2r - 1$ and $b_3 > b_2$. Then for $x \in (b_2, b_3)$, we have $\inf_{n\rightarrow\infty} g_0(x) \geq 0$.

(3) Note that $C(x)$ increases and $D(x)$ decreases, then via a similar process, we find that for arbitrarily small positive number ϵ, when $x \in ((a_0+\epsilon)/k, b_2-\epsilon)$, there exists $\delta = max(\frac{1}{2}g_1(a_0+\epsilon), g_2(b_2-\epsilon))$ such that $\sup_{n\rightarrow\infty} g_0(x) < -\delta$.

4.4. Case D where $d \rightarrow \infty$, $k \rightarrow \infty$, $k = bn$, and $0 < b < 1$

(1) If $x = a/n$, where $a = const$, we have $\sigma(x) \rightarrow (1-b)^a$,

$$g_0(x) \rightarrow 1 - r + r\frac{1}{-\ln(1-p)} \ln\left(1-p+p(1-b)^a\right) \triangleq g_3(a).$$

Solving equation $g_3(a) = 0$ with respect to variable a, solution a_1 is obtained. Similar to Case C, if $x \in (0, a_1/n)$ we have $\inf_{n\rightarrow\infty} g_0(x) \geq 0$.

(2) As the same with Case C, if x is a positive constant, $g_0(x) \rightarrow g_2(x)$; then for b_2 and b_3, we have $\inf_{n\rightarrow\infty} g_0(x) \geq 0$ for $x \in (b_2, b_3)$.

(3) Similar to Case C, for arbitrarily small positive number ϵ, when $x \in ((a_1+\epsilon)/n, b_2-\epsilon)$, there exists $\delta = \max\{\frac{1}{2}g_3(a_1+\epsilon), g_2(b_2-\epsilon)\}$ such that $\sup_{n\rightarrow\infty} g_0(x) < -\delta$.

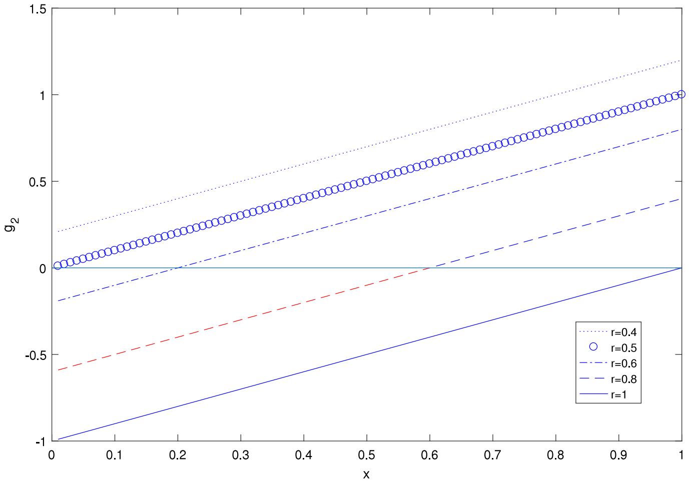

Fig. 3. Function $g_2(x)$, where $r = 0.4, 0.5, 0.6, 0.8, 1$ from top to bottom. There is a region satisfying $g_2(x) < 0$ when $r > 0.5$. Taking $r = 0.8$ for example, in the region $x \in (0, 0.6)$, $g_2(x) < 0$ (plotted in red).

Table 2
Details of clustering phenomenon in different cases.

Cases	Clustering range	Cluster diameter $\leq$	Distance among cluster-regions $\geq$	Number of clusters
Case A	$0.5 < r < 1$	$a_0 n/k$	$b_0 n - a_0 n/k \approx b_0 n$	Exponential
Case B	$0.5 < r < 1$	a_1	$b_0 n - a_1 \approx b_0 n$	Exponential
Case C	$0.5 < r < 1$	$a_0 n/k$	$(2r-1)n - a_0 n/k \approx (2r-1)n$	Exponential
Case D	$0.5 < r < 1$	a_1	$(2r-1)n - a_1 \approx (2r-1)n$	Exponential
Model RB [26]	$r_0 < r < 1$	$a_2 n$	$(b_2 - a_2)n$	Exponential

5. The number of clusters

Let $l = \max_{0<xn\leq \alpha n} \mathbb{E}(Z(x))$, where $\alpha = a_0/k$ in Cases A and C and $\alpha = a_1/n$ in Cases B and D. Therefore, for $0 < x \leq \alpha$, in cases A, B, C, and D, we have $x \to 0$ and

$$\ln\left(\mathbb{E}(Z(x))\right)/(n \ln d) \to D(x) = 1 - r + r\frac{1}{-\ln(1-p)} \ln\left(1 - p + p\sigma(x)\right) \leq 1 - r,$$

where the last inequality is because $0 \leq \sigma(x) \leq 1$. By the above inequality, we obtain

$$\sup_{n\to\infty} l \leq n^{\ln d - r \ln d}.$$

When $r < 1$, we obtain a lower bound of the value of Eq. (2), and the number of cluster-regions is larger than

$$\frac{\frac{1}{n} d^n (1-p)^t}{n l^{0.5}} \geq \frac{1}{n^2} d^{0.5(1-r)n}.$$

Thus the number of cluster-regions increases exponentially with n.

6. Clustering phase of model *d*-*k*-CSP

The method in Section 3 implies that, if solution-pairs at distance between αn and βn do not exist, then the clustering phase appears where the cluster diameters are smaller than αn and the distance among clusters are larger than $\beta n - \alpha n$. In Section 4 we found such α, β for Cases A, B, C and D and $0.5 < r < 1$. And in Section 5 we obtained the number of clusters, which increases exponentially with n. We list all those properties in Table 2.

In Table 2, $a_0 = -\ln \frac{(1-p)^{1/r} - (1-p)^2}{1-p-(1-p)^2}$; $a_1 = \frac{\ln((1-p)^{\frac{1-r}{r}} - 1 + p) - \ln p}{\ln(1-b)}$; $b_0 \in (0, \frac{d-1}{d})$ is the solution of the function $\ln d - \ln(x^x(1-x)^{1-x}) + x\ln(d-1) - 2r\ln d = 0$ with respect to variable x. a_2 and b_2 are the two solutions of equation $\alpha(1+x) + r\ln[(1-p)^2 + p(1-p)(1-x)^k] = 0$ with respect to variable x; and r_0 is the smallest value for which this equation has not least a solution in $x \in [0, 1]$. The results of model RB are from Ref. [26]. We have the following observations.

(1) There is an exponential number of cluster-regions, with each cluster-region containing a sub-exponential number of solutions.

Table 3
Details of clustering phenomenon in the special case where $k = n$.

	Clustering range	Cluster diameter	Distance among clusters	Number of clusters
$k = n$ and $d = const$	$0.5 < r < 1$	0	$b_0 n$	Exponential
$k = n$ and $d \to \infty$	$0.5 < r < 1$	0	$(2r - 1)n$	Exponential

(2) The smallest distances among cluster-regions are $\Theta(n)$ (of the same order of n), so the cluster-regions are well-separated.
(3) With r approaching 1, the diameter of a cluster-region decreases to a small value.
(4) With r approaching 1, the distance among the cluster-regions increases, which also means that the number of cluster-regions decreases because of that some cluster-regions disappear.
(5) For fixed r, as k increases, the diameter of the cluster-region decreases. When k is a constant, the biggest diameter is $a_2 n$; when $k \to \infty$ but $k = o(n)$, the biggest diameter is $a_0 n/k$; when $k = bn$, with $0 < b < 1$, it is a_1. In summary the biggest diameter is $\Theta(n/k)$.
(6) For fixed r, as d increases, the distance among the cluster-regions increases. When d is a constant in case A, B, the smallest distance is $b_0 n$; when $d \to \infty$ in Cases C and D, the smallest distance is $(2r - 1)n$.
(7) Condensation phase does not exist, because when $0.5 < r < 1$ it is in clustering phase and when $r > 1$ it is in unsatisfiable phase.

From the above observations (1–3), when r is below and close to 1, the solution space contains many well-separated small cluster-regions.

7. A special case where $k = n$

If $k = n$, for all $xn = 1, 2, \ldots, n$, we have $\sigma(x) = 0$, and then from Eq. (3) we have

$$\mathbb{E}(Z(x)) \to d^n C_n^{nx} (d-1)^{nx} (1-p)^{2t}.$$

Furthermore if $d = const$, we have $f_0(x) \to f_2(x)$ for $x = 1/n, 2/n, \ldots, 1$, where $f_2(x)$ is defined in Eq. (4) and is shown in Fig. 2. If $d \to \infty$, we have $g_0(x) \to g_2(x)$ for $x = 1/n, 2/n, \ldots, 1$, where $g_2(x)$ is defined in Eq. (5) and is shown in Fig. 3. Therefore, when $0.5 < r < 1$ and $d = const$, w.h.p. solution-pairs at a distance between 1 and $b_0 n$ do not exist; when $0.5 < r < 1$ and $d \to \infty$ w.h.p. solution-pairs at any distance between 1 and $2r - 1$ do not exist. Then, this special case has a special solution space structure that the solutions are isolated from each other. If we regard an isolated solution as a cluster, the details of the clustering can be shown in Table 3. Notice that the cluster diameter is 0, which means that each cluster only contains a single solution.

8. Marginals of variables and distribution of clusters

To find out the marginals of variables and the distribution of clusters, we make use of a physical method which is called Bethe–Peierls approximation or Belief Propagation (BP). The Belief Propagation is an iterative message-passing algorithm, where the update rules are that

$$\nu_{i\to a}^{(t+1)}(x_i) \cong \prod_{b\in\partial i\setminus a} \widehat{\nu}_{b\to i}^{(t)}(x_i), \tag{6}$$

$$\widehat{\nu}_{a\to i}^{(t)}(x_i) \cong \sum_{\underline{x}_{\partial a\setminus i}} \psi_a(\underline{x}_{\partial a}) \prod_{k\in\partial a\setminus i} \nu_{k\to a}^{(t)}(x_k), \tag{7}$$

where $\nu_{i\to a}^{(t)}$, $\widehat{\nu}_{a\to i}^{(t)}$ are the messages between variable node i and its adjacent constraint node a at step t. The symbol $\cong$ denotes equality up to a normalization, because BP messages are understood to be probability distributions. ψ_a is equal to 1 if constraint a is satisfied by $\underline{x}_{\partial a}$, and is equal to 0 otherwise. After the iteration converging, let $\nu_{i\to a}^{(t)}$ converge to $\nu_{i\to a}$ and $\widehat{\nu}_{a\to i}^{(t)}$ converge to $\widehat{\nu}_{a\to i}$, then the marginal probability that variable i equals to x_i is

$$\nu_i(x_i) \cong \prod_{b\in\partial i} \widehat{\nu}_{b\to i}(x_i). \tag{8}$$

Firstly, we study on whether the BP iteration converges. To do experiments, we should generate typical d-k-CSP instances. Here we let $n = 200$, and when $d = 2$ we let $k \geq 8$; when $d = 3$ let $k \geq 5$; when $k = 2$ let $d \geq 15$. In Figs. 4 and 5, we show probabilities of convergence for different ks and ds. It shows that the BP iteration converges when k or d is large enough. Actually, starting from uniformly random messages, it converges to a unique fixed point.

When both k and d are small, the BP iteration tends to be non-convergent, which is reasonable because of the solution space structure. When both k and d are small, the problem tends to be unsatisfiable, which can be supported by the

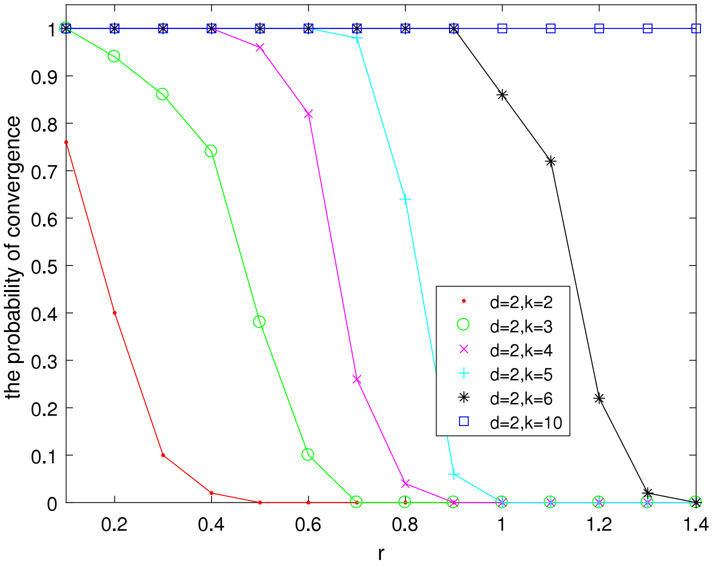

Fig. 4. Probabilities of convergence of BP iteration, where we set $n = 200, d = 2, k = 2, 3, 4, 5, 6, 10, p = 0.5, r = 0.1, \ldots, 1.4$. Each point is averaged over 50 instances.

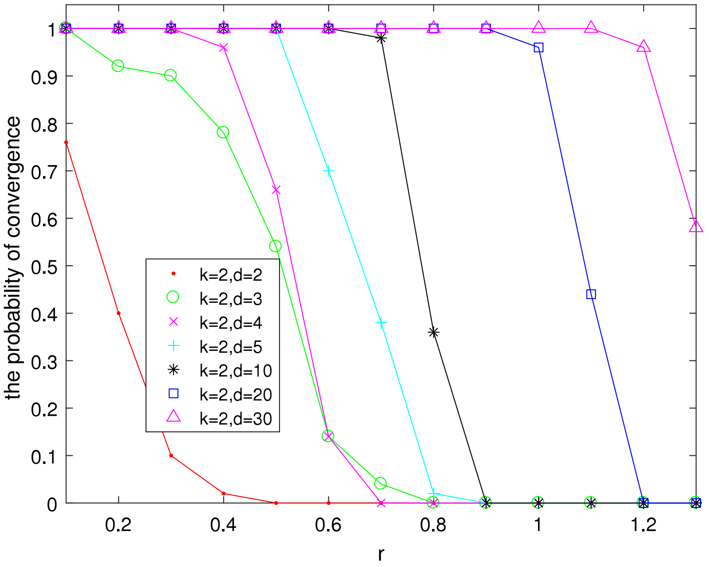

Fig. 5. Probabilities of convergence of BP iteration, where we set $n = 200, k = 2, d = 2, 3, 4, 5, 10, 20, 30, p = 0.5, r = 0.1, \ldots, 1.3$. Each point is averaged over 50 instances.

performance of local search algorithm. When k or d is large enough, the instance is a typical instance, and the instance is unsatisfiable for $r > 1$, but experiments show that the BP iteration still converges. For example, when $k = 10$ in Fig. 4, the BP iteration converges until $r = 2.3$, which is much bigger than $r = 1$. We believe the reasons for the convergence when $r > 1$ are that: (1) Constraint function $\psi_a(\underline{x})$ rarely forbids variable $i \in \partial a$ from being a specific value x_i, so only positive numbers are propagating; (2) The factor graph is over-connected, so a variable receives a lot of different messages but after calculation it always has to treat each of the values equally. Although the solution space is empty, the iteration still converges.

Secondly, we study on whether BP gives the correct marginals. Bethe free entropy is a function of converged messages $\nu_{i\to a}$ and $\widehat{\nu}_{a\to i}$,

$$S_{Bethe} = \sum_a S_a + \sum_i S_i - \sum_{(i,a)} S_{ia},$$

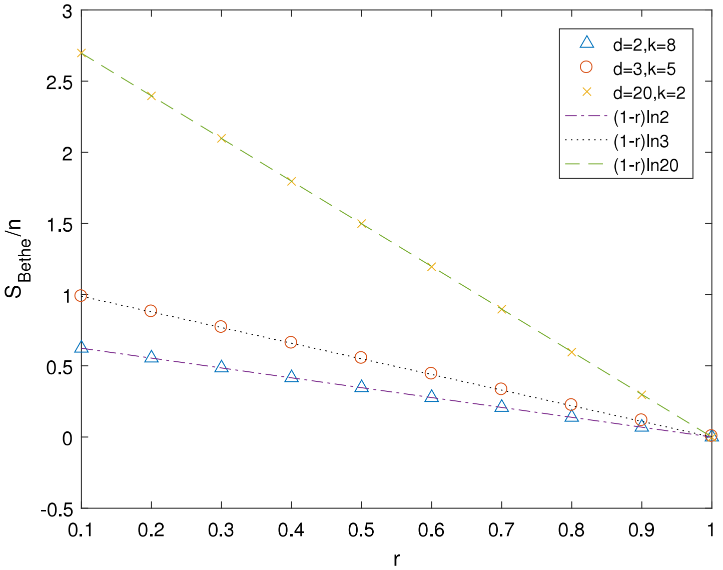

Fig. 6. Bethe free entropy S_{Bethe}/n, where $r = 0.1, 0.2, \ldots, 1$, $n = 200$, $p = 0.5$. It shows that S_{Bethe}/n coincides with $\ln[\mathbb{E}(X)]/n = (1-r)\ln d$. Each point is averaged over 30 instances.

where

$$S_a = \log\left[\sum_{\underline{x}_{\partial a}} \psi_a(\underline{x}_{\partial a}) \prod_{i\in\partial a} \nu_{i\to a}(x_i)\right],$$

$$S_i = \log\left[\sum_{x_i} \prod_{b\in\partial i} \widehat{\nu}_{b\to i}(x_i)\right],$$

$$S_{ia} = \log\left[\sum_{x_i} \nu_{i\to a}(x_i)\widehat{\nu}_{a\to i}(x_i)\right].$$

Here we take three examples, and calculate the Bethe free entropy. In Fig. 6, it shows that S_{Bethe}/n coincides with the logarithm of the average number of solutions divided by n (annealed entropy density), which is $\ln[\mathbb{E}(X)]/n = (1-r)\ln d$. This indicates that the replica symmetry solution should always be stable locally, and that BP gives the correct marginals.

Experiments in the next section show that BP guided algorithms can find solutions before $r = 0.9$. Both the research on Bethe free entropy and BP guided algorithms suggest that BP gives the correct marginals.

Thirdly, we study on the marginals of variables and the distribution of clusters. For $d = 2$, because there are only two marginals for each variable, we can research on the marginal difference, which equals to the bigger marginal minus the smaller marginal. Fig. 7 is obtained by Belief Propagation, which shows that marginal difference is small for $k = 10$, and with the k increasing, the difference decreases. It shows that the marginals are very uniform when k is big.

When d is large, we take an example and estimate the marginals by Belief Propagation. We set $n = 200, d = 15, k = 2, p = 0.5, r = 0.9$, and the result in Fig. 8 shows that almost all marginals are positive and many of them are around the average value 0.05.

To sum up, the examinations show that the marginals are uniform to a certain extent, which means that the clusters distribute widely in the solution space. Our claim of "widely distributed clusters" is based on an elementary idea that, if the total number of solutions are fixed, the more widely the solutions distribute, the harder it is to determine their positions. In the d-k-CSP model we study, because the clusters distribute widely, although BP gives the correct marginals of variables, the marginals are almost uniform. This may trap algorithms guided by marginals. After dynamical transition, if clusters do not reduce the size in the same speed, dominant or subdominant [33] clusters may appear and expose its position. Uniform distribution also should be the reason that the satisfiable–unsatisfiable transition point can be determined by the first moment method and the second moment method. In a random model, where constraints do not have a specific structure, the existence of widely distributed clusters may be one of the reasons for the difficulty of the problem.

9. Whitening procedure and frozen variables

Firstly, we find solutions by BP decimation algorithm (BPD) [34] and reinforced BP algorithm (RBP) [35].

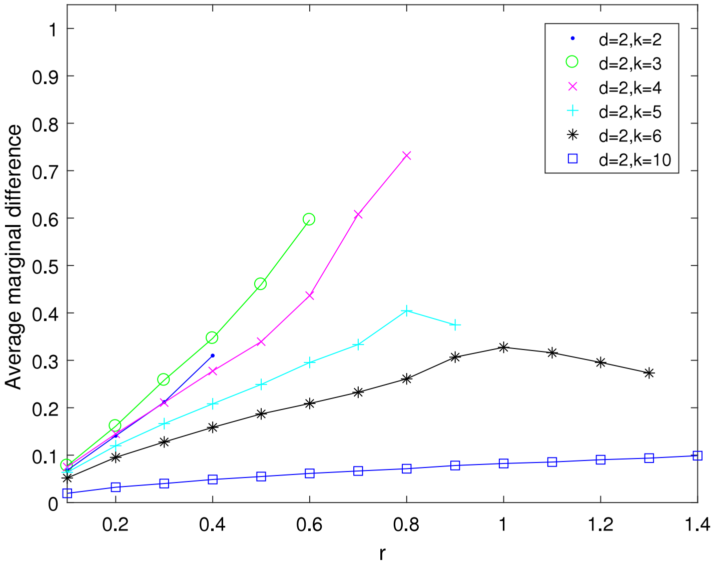

Fig. 7. Average marginal differences, where we set $n = 200, d = 2, k = 2, 3, 4, 5, 6, 10, p = 0.5, r = 0.1, \ldots, 1.4$. Each point is averaged over 50 instances.

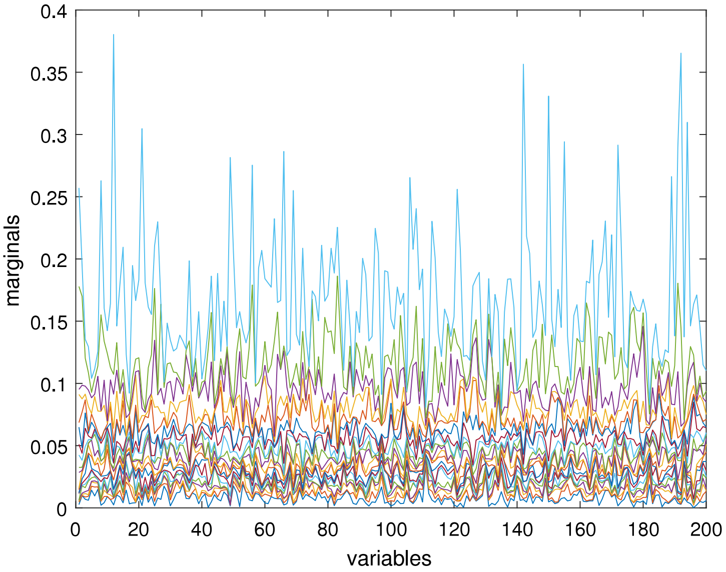

Fig. 8. Marginals of 200 variables where we set $n = 200, d = 20, k = 2, p = 0.5, r = 0.9$. The top line represents the biggest marginal probability for each variable; the second line from above represents the second biggest marginal probability for each variable; and so on. The averages of the values in different lines are 0.1681 0.1168 0.0954 0.0805 0.0711 0.0628 0.0555 0.0497 0.0446 0.0405 0.0362 0.0323 0.0287 0.0255 0.0226 0.0197 0.0171 0.0146 0.0112 0.0074.

BP decimation algorithm performs BP iteration, where the update rules are (6) and (7). After BP iteration converges, the algorithm will estimate marginals by (8); find the most polarized variable (the one which has the largest marginal probability); fix the variable to its most possible value; and reduce the problem. Repeating those steps n times, if BP iteration always converges, we obtain an assignment of the n variables. If the assignment is a solution, the solving process is successful, otherwise it fails.

Reinforced BP algorithm in each iteration step follows update rules (6) and (7) with probability p, and follows update rules (7), (9) and (10) with probability $1 - p$.

$$\nu_{i\to a}^{(t+1)}(x_i) \cong \mu_i^{(t)}(x_i) \prod_{b\in\partial i\setminus a} \widehat{\nu}_{b\to i}^{(t)}(x_i), \tag{9}$$

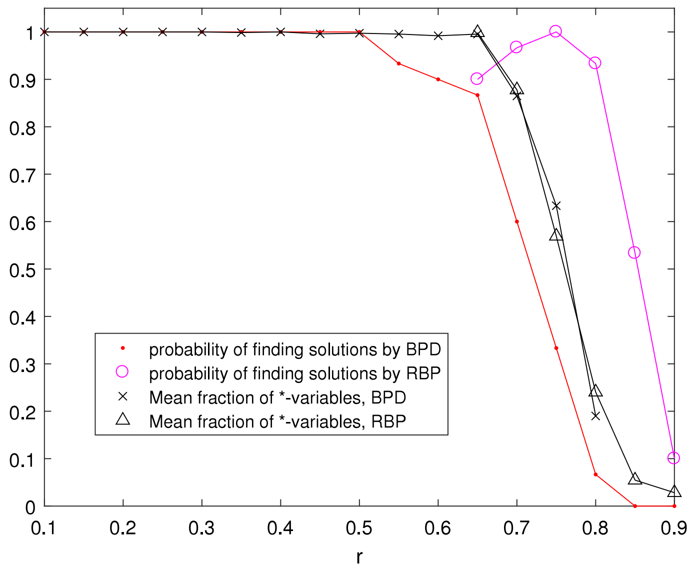

Fig. 9. Performances of BPD, RBP, and whitening procedure, where $r = 0.1, 0.15, \ldots, 0.9$, $n = 200$, $p = 0.5$, $d = 2$, $k = 8$. The average is computed over 30 instances for each r.

$$\mu_i^{(t)}(x_i) \cong \prod_{b\in\partial i} \widehat{\nu}_{b\to i}^{(t-1)}(x_i). \tag{10}$$

The iteration does not need to converge, but in each iteration step the algorithm estimates marginals by

$$\nu_i(x_i) \cong \prod_{b\in\partial i} \widehat{\nu}_{b\to i}^{(t)}(x_i),$$

and assigns every variable to its most possible value, and verifies whether the assignment is a solution. If the maximum number of iteration steps has been reached and still no solutions are found, algorithm fails. Here we set the maximum number of iteration steps to be 1000, set p to be $t^{-0.1}$.

We use the algorithms on the same instances as in Fig. 6. Result shows that BPD solves the problem with probability 1 for small r, and find solutions with positive probability before $r = 0.8$. Figs. 9–11 show that RBP can find solutions with positive probability in the region $0.6 < r < 0.9$.

Secondly, we perform the whitening procedure [14,36] on the solutions found by BPD and RBP. For any solution, every variable has one value. If variables can have a set of values, it actually composes a region of assignments. The whitening procedure expands a solution to a region of assignments, and the region contains the cluster to which the solution belongs. In the first step, starting from a solution, if a variable can be reassigned other values without violating any constraints, expand the value of the variable to the value-set (the set of the values). In subsequent steps, expand the value (or the value-set) of a variable, if there are other values of the variable that can constitute a solution with some values of adjacent variables from their value-set. Since this is an expanding procedure, the procedure will terminate eventually. After the procedure reaches a fixed point, we have a value-set for every variable. If the value-set has only one value, the variable is frozen; if the value-set has d values, the variable is $*-$variable.

Fig. 9 shows that whitening procedures on solutions found by BPD and RBP have the same behavior. In Fig. 9, d equals to 2, so the fraction of $*-$variables equals to the fraction of non-frozen variables, which drops fast in the region $0.65 < r < 0.85$. When $r > 0.85$, more than 90 percent of the variables are frozen. In Figs. 10 and 11, d equals to 3 and 20 respectively, so the fraction of non-frozen variables is bigger than the fraction of $*-$variables, both of which drop in the region $0.7 < r < 0.9$. When $r > 0.9$, more than 75 percent of the variables are frozen.

To sum up, experiments show that variables become frozen before $r = 1$. The fact that variables become frozen may be the reason that RBP algorithm fails, because it shows that when the number of frozen variables increase before $r = 0.9$, the RBP becomes ineffective. When all variables are frozen, where $0.9 < r < 1$, RBP cannot find solutions. In the region $0.7 < r < 0.9$, there also should be a rigidity transition, where typical solutions are frozen and a minority of solutions are still unfrozen. The rigidity transition and freezing transition should be further studied for the model, in order to make the picture of transitions more clearly.

10. Conclusion

Model d-k-CSP is a standard prototype of Constraint Satisfaction Problem (CSP), where the domain size d and/or the length of constraint scope k grow with the number of variables n. Firstly, we use a mathematical method to show that,

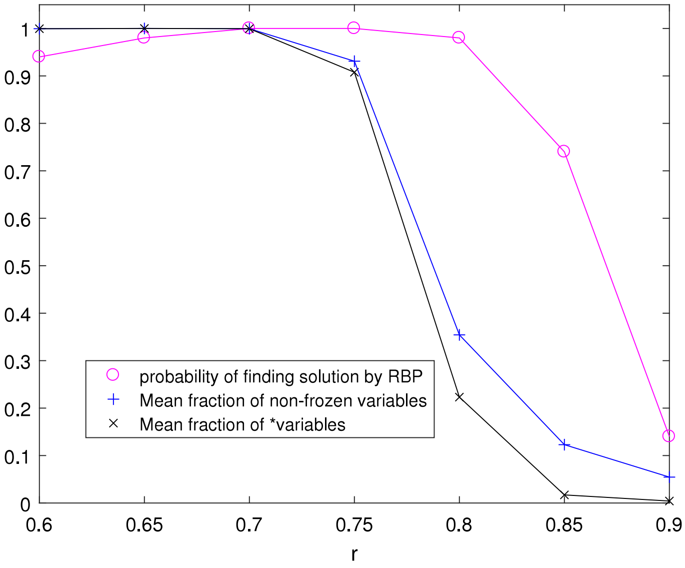

Fig. 10. Performances of RBP and whitening procedure, where $r = 0.6, 0.65, \ldots, 0.9$, $n = 200$, $p = 0.5$, $d = 3$, $k = 5$. The average is computed over 50 instances for each r.

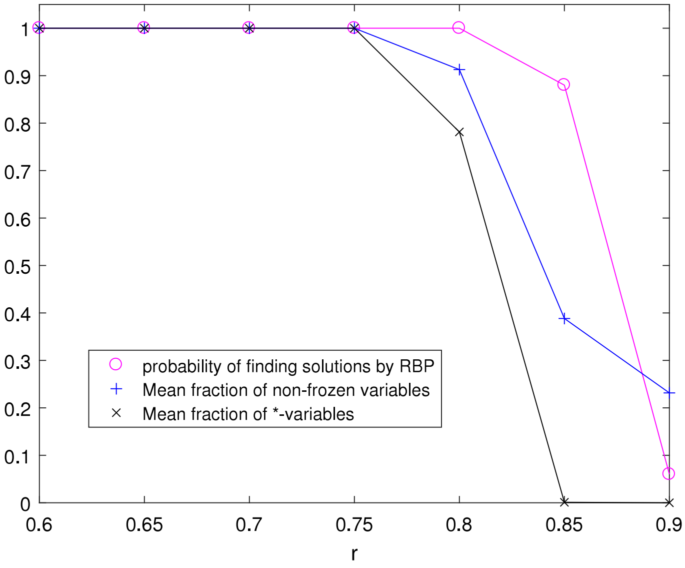

Fig. 11. Performances of BPD, RBP, and whitening procedure, where $r = 0.6, 0.65, \ldots, 0.9$, $n = 200$, $p = 0.5$, $d = 20$, $k = 2$. The average is computed over 50 instances for each r.

before the satisfiable–unsatisfiable transition, the solution space shatters into an exponential number of well-separated cluster-regions, and the diameter of a cluster decreases with r. Secondly, physical method shows that the clusters distribute widely in the solution space. Thirdly, whitening procedure on solutions shows that variables become frozen before $r = 1$. So when r is below and close to 1, the solution space contains many widely distributed and well-separated small frozen clusters. When r is below and close to 1, the instances are hard to solve, so this solution space structure will lead to highly hard problems.

Also we consider the d-k-CSP model from several cases, and give the properties of the clustering phenomenon, including cluster-region diameter, distance between cluster-regions, clustering range, and number of the clusters. For fixed r, cluster-region diameter depends on k; distance between cluster-regions depends on d. As k increases, the diameter of the cluster-region decreases, and the diameter is smaller than $\Theta(n/k)$. As d increases, the distance among the cluster-regions increases.

Acknowledgments

Project supported by the Fundamental Research Funds for the Central Universities (No. FRF-TP-16-065A1), and the National Natural Science Foundation of China (No. 11801028 and No. 61702019).

References

[1] P. Cheeseman, B. Kanefsky, W.M. Taylor, Proc. IJCAI, 1991, pp. 331-337.
[2] E. Friedgut, J. Bourgain, J. Amer. Math. Soc. 12 (1999).
[3] M. Mézard, G. Parisi, Eur. Phys. J. B 20 (2001) 217–233.
[4] M. Mézard, R. Zecchina, Phys. Rev. E 66 (2002) 056126.
[5] M. Mézard, G. Parisi, R. Zecchina, Science 297 (2002) 812.
[6] M. Mézard, F. Ricci-Tersenghi, R. Zecchina, J. Stat. Phys. 111 (2003) 505–533.
[7] F. Krzakala, A. Montanari, F. Ricci-Tersenghi, G. Semerjian, L. Zdeborova, Proc. Natl. Acad. Sci. USA 104, 2007, pp. 10318-10323.
[8] A. Montanari, F. Ricci-Tersenghi, G. Semerjian, J. Stat. Mech. Theory Exp. (2008) P04004.
[9] L. Zdeborova, F. Krzakala, Phys. Rev. E 76 (2007) 031131.
[10] A. Braunstein, M. Mézard, R. Zecchina, Rand. Struct. Algorithms 27 (2002) 201–226.
[11] M. Angelini, F. Ricci-Tersenghi, arXiv: Disordered Systems and Neural Networks, 2019.
[12] S. Seitz, M. Alava, P. Orponen, J. Stat. Mech. Theory Exp. 2005 (06) (2005) 06006.
[13] J. Ardelius, E. Aurell, Phys. Rev. Lett. 74 (3) (2006) 037702.
[14] R. Marino, G. Parisi, F. Ricci-Tersenghi, Nature Commun. 7 (2016) 12996.
[15] I.P. Gent, E. Macintyre, P. Prosser, B.M. Smith, T. Walsh, Constraints 6 (2001) 345–372.
[16] B. Smith, M. Dyer, Artificial Intelligence 81 (1996) 155–181.
[17] D. Achlioptas, L.M. Kirousis, E. Kranakis, M. Molloy, Y.C. Stamatiou, Proc. Principles and Practice of Constraint Programming, 1997, pp. 107-120.
[18] Y. Gao, J. Culberson, J. Artificial Intelligence Res. 28 (2007) 517–557.
[19] B. Smith, Theoret. Comput. Sci. 265 (2001) 265–283.
[20] A. Frieze, M. Molloy, Rand. Struct. Algorithms 28 (2006) 323–339.
[21] K. Xu, W. Li, J. Artificial Intelligence Res. 12 (2000) 93–103.
[22] Y. Fan, J. Shen, Artificial Intelligence 175 (2011) 914–927.
[23] Y. Fan, J. Shen, K. Xu, Artificial Intelligence 193 (1–17) (2012) 1017–1054.
[24] K. Xu, F. Boussemart, F. Hemery, C. Lecoutre, Artificial Intelligence 171 (8) (2007) 514–534.
[25] C. Zhao, P. Zhang, Z. Zheng, K. Xu, Phys. Rev. E 85 (2012) 016106.
[26] W. Xu, P. Zhang, T. Liu, F. Gong, J. Stat. Mech. Theory Exp. (2015) P12006.
[27] M. Mézard, T. Mora, R. Zecchina, Phys. Rev. Lett. 94 (2005) 197205.
[28] D. Achlioptas, F. Ricci-Tersenghi, Proc. STOC'06, 2006, pp. 130-139.
[29] D. Achlioptas, A. Coja-Oghlan, F. Ricci-Tersenghi, Rand. Struct. Algorithms 38 (2011) 251–268.
[30] D. Achlioptas, Eur. Phys. J. B 64 (2008) 395–402.
[31] D. Achlioptas, A. Coja-oghlan, Proc. Graph-theoretic Concepts in Computer Science, 2010.
[32] L. Zdeborova, M. Mézard, Phys. Rev. Lett. 101 (7) (2008) 078702.
[33] C. Baldassi, A. Ingrosso, C. Lucibello, Phys. Rev. Lett. 115 (12) (2015) 128101.
[34] F. Ricci-Tersenghi, G. Semerjian, J. Stat. Mech. Theory Exp. 2009 (09) (2009) 09001.
[35] A. Braunstein, Z. Riccardo, Phys. Rev. Lett. 96 (3) (2006) 030201.
[36] G. Parisi, J. Math. Phys. 49 (2008) 125216.