Reconstruction and Clustering in Random Constraint Satisfaction Problems

Andrea Montanari* Ricardo Restrepo† Prasad Tetali‡

April 17, 2009

Abstract

Random instances of Constraint Satisfaction Problems (CSP's) appear to be hard for all known algorithms, when the number of constraints per variable lies in a certain interval. Contributing to the general understanding of the structure of the solution space of a CSP in the satisfiable regime, we formulate a set of natural technical conditions on a large family of (random) CSP's, and prove bounds on three most interesting thresholds for the density of such an ensemble: namely, the *satisfiability* threshold, the threshold for *clustering* of the solution space, and the threshold for an appropriate *reconstruction* problem on the CSP's. The bounds become asymptoticlally tight as the number of degrees of freedom in each clause diverges. The families are general enough to include commonly studied problems such as, random instances of Not-All-Equal-SAT, k-XOR formulae, hypergraph 2-coloring, and graph k-coloring. An important new ingredient is a condition involving the Fourier expansion of clauses, which characterizes the class of problems with a similar threshold structure.

* Departments of Electrical Engineering and Statistics, Stanford University; research funded in part by the NSF grants CCF-0743978 and DMS-0806211

† School of Mathematics, Georgia Tech, Atlanta, GA 30332-0160

‡ Schools of Mathematics and Computer Science, Georgia Tech, Atlanta, GA 30332-0160; research funded in part by the NSF grant DMS-0701043

1 Introduction

Given a set of n variables taking values in a finite alphabet, and a collection of m constraints, each restricting a subset of variables, a Constraint Satisfaction Problem (CSP) requires finding an assignment to the variables that satisfies the given constraints. Important examples include k-SAT, Not All Equal SAT, graph (vertex) coloring with k colors etc. Understanding the threshold of satisfiability/unsatisfiability for *random* instances of CSPs, as the number of constraints $m = m(n)$ varies, has been a challenging task for the past couple of decades, with some notable successes (see e.g., [ANP05]). On the algorithmic side, the challenge of *finding* solutions of a random CSP *close to the threshold of satisfiability* (in the regime where solutions are known to exist) remains widely open. All provably polynomial-time algorithms fail well before the SAT to UNSAT threshold.

The attempt to understand this universal failure led to studying the geometry of the set of solutions of random CSPs [MPZ02, AC08], as well as the emergence of long range correlations among variables in random satisfying assignments [KM+07]. These research directions are motivated by two heuristic explanations of the failure of polynomial algorithms: **(1)** The space of solutions becomes increasingly complicated as the number of constraints increases and is not captured correctly by simple algorithms; **(2)** Typical solutions become increasingly correlated and local algorithms cannot unveil such correlations.

By analyzing a large class of random CSP ensembles, this paper provides strong support to the belief that the above phenomena are *generic*, that they are characterized by *sharp thresholds*, and that the *thresholds for clustering and reconstruction do coincide*.

1.1 Related work

Building on a fascinating conjecture on the geometry of the set of solutions, statistical physicists have developed surprisingly efficient message passing algorithms to solve random CSPs. For instance, survey propagation [MPZ02, MZ02] has been shown empirically to find solutions of random 3-SAT extremely close to the SAT-UNSAT transition. In order to understand the success of these heuristics, it has become important to study the thresholds for the emergence of so-called *clustering* of solutions – the emergence of an exponential number of sets (or clusters) of solutions, where solutions within a cluster are closer (in the Hamming sense, say), compared to the intra-cluster distance [MMZ05, AR06, AC08]. Moreover, the fact that solutions within a cluster impose long-range correlations among assignments of variables, motivates one to study the so-called reconstruction problem in the context of random CSP's. Indeed, non-rigorous statistical mechanics calculations imply that the clustering and reconstruction thresholds coincide [MM06, KM+07].

Finally, understanding the threshold for (non)reconstruction is also becoming relevant (if not crucial) to understanding the limit of the Glauber dynamics to sample from the set of solutions of a CSP. Indeed non-reconstuctibility was proved in [BK+05] to be a necessary condition for fast mixing, and is expected to be sufficient for a large class of 'sufficiently random' problems [GM07].

In a recent paper, Gerschenfeld and the first author [GM07], considered the reconstruction problem for *graphical models*, which included the case of proper colorings of the vertices of a random graph. This amounts to understanding the correlation (as measured e.g. through mutual information) between the color of a vertex v, and the colors of vertices at distance $\geq t$ from v. In particular, the problem is said to be 'unsolvable' if such a correlation decays to 0 with t. We refer to Section 3 for a precise definition of the reconstruction problem. For a class of models, including the so-called Ising spin glass, the antiferromagnetic Potts model, and proper q-colorings of a graph, [GM07] derived a general sufficient condition, under which reconstruction for (sparse) random graphs $G(n, m)$ with $m = cn$ edges is possible if and only if it is possible for a Galton-Watson tree with independent Poisson($2c$) degrees for each vertex. Moreover, they also verified that the condition holds for the Ising spin glass and the antiferromagnetic Potts at non-zero temperature, leaving open the case of proper colorings of graphs.

1.2 Summary of contributions

It is against this backdrop that we consider certain general families of CSP's – the first dealing with constraints consisting of k-tuples of binary variables (as in k-uniform hypergraph 2-coloring or Not-All-Equal (NAE) k-sat), while the second dealing with q-colorings of vertices of graphs (which may be seen as an instance of a CSP with q-ary variables) – and study three important threshold phenomena. Our chief contribution is as follows.

(a) We formulate a fairly natural set of assumptions under which a general class of constraint satisfaction problems (including the models mentioned above) can be understood rather precisely in terms of the thresholds for satisfiability, clustering and (non)reconstruction phenomena. In particular we verify that the last two thresholds coincide within the precision of our bounds.

(b) We consider tree ensembles (families of random CSP's whose variable-constraint dependency structure takes the form of a tree), and prove optimal bounds on the threshold for reconstruction on trees. These CSP's consist of binary variables, and the constraints are k-ary, and the bounds are optimal to first order, as k goes to infinity.

(c) We verify the sufficient condition of [GM07] for proper colorings of graphs, thus extending the reconstruction result for colorings on trees to the same on (sparse) random graphs.

(d) By way of techniques, we make crucial use of the Fourier expansion of the (binary k-CSP) constraints, after introducing an assumption on the Fourier expansion, as part of the random ensemble under consideration; this is key to being able to characterize the thresholds precisely.

(e) Finally, as illustrative examples, we mention the specific bounds (on various thresholds) that follow for some standard models, such as the NAE k-SAT, k-XOR formulae etc.

The organization of the paper is as follows. In Section 2, we give the formal definitions and assumptions of our models. We state our main results in Section 3. In Section 4, we state and prove the optimal bounds for the tree reconstruction problem. In Section 5, we verify the sufficient condition (from [GM07]) for the specific problem of graph proper q-coloring, thus proving one of our main results – optimal bounds on the (sparse) random graph reconstruction problem for colorings. In Appendix A, we derive a certain technical second moment bound that is needed for our work.

2 Definitions

In this section we define a family of random CSP ensembles: problems with constraints involving k-tuples of binary variables and q-ary ensembles as a natural extension. We also introduce some analytic definitions that we will need in order to present our results.

Binary k-CSP ensemble. Given an integer n, $\alpha \in \mathbb{R}_+$, and a distribution $p = \{p(\varphi)\}$ over Boolean functions $\varphi : \{+1, -1\}^k \to \{0, 1\}$, $\text{CSP}(n, \alpha, p)$ is the ensemble of random CSP's over n Boolean variables $\underline{x} = (x_1, \ldots, x_n)$ defined as follows. For each $a \in \{1, \ldots, m = n\alpha\}$, draw k indices $i_a(1), \ldots, i_a(k)$ independently and uniformly at random in $[n]$, and a function φ_a with distribution $p(\varphi)$. An assignment $\underline{x}$ satisfies the resulting instance if $\varphi_a(x_{i_a(1)}, \ldots, x_{i_a(k)}) = 1$ for each $a \in [m]$. A CSP instance can be naturally described by a bipartite graph G (often referred to in the literature as a 'factor graph') including a node for each clause $a \in [m]$ and for each variable $i \in [n]$, and an edge (i, a) whenever variable x_i appears in the a-th clause.

q-ary ensembles. A q-ary ensemble is the natural generalization of a binary ensemble to the case in which variables take q values. For the sake of simplicity, we restrict our discussion here to the case of pairwise constraints (i.e. $k = 2$ in the language of the previous section).

Given an integer n, $\alpha \in \mathbb{R}_+$, and a distribution $p = \{p(\varphi)\}$ over Boolean functions $\varphi : [q] \times [q] \to \{0, 1\}$, $\text{CSP}_q(n, \alpha, p)$ is the collection of random CSP's over q-ary variables x_i, for $i = 1, 2, \ldots, n$, defined as follows. For each $a \in \{1, \ldots, m = n\alpha\}$, draw 2 indices i_a, j_a independently and uniformly at random in $[n]$, and a function φ_a with distribution $p(\varphi)$. An assignment $\underline{x} = (x_1, \ldots, x_n)$ satisfies the resulting instance, if $\varphi_a(x_{i_a}, x_{j_a}) = 1$ for each $a \in [m]$.

In this paper, by way of illustrating how the results for binary ensembles could be (purportedly) extended to q-ary ensembles, we will exclusively study the q-coloring model which consists of ensembles with the single clause $\varphi(x, y) = \mathbb{I}(x \neq y)$. This model corresponds to proper colorings with q colors of a random sparse graph with an edge-to-vertex density of $\alpha > 0$.

In the rest of this section, we briefly review some well known definitions in discrete Fourier analysis that are useful for stating our results.

Functional analysis of clauses. We denote by v_θ, the measure defined over $\{-1,+1\}^k$ such that $v_\theta(x) = \prod_{i=1}^{k}\left(\frac{1+x_i\theta}{2}\right)$ for every $x \in \{-1,+1\}^k$. This is just the measure induced by choosing k independent copies of a random variable that takes values ± 1 and has expectation θ. Notice that when $\theta = 0$, v_θ corresponds to the uniform measure over $\{-1,+1\}^k$.

The inner product induced by this measure, on the space of real functions defined on $\{-1,+1\}^k$ is denoted by $(\cdot,\cdot)_\theta$, and the correponding norm by $\|\cdot\|_\theta$. If $\theta = 0$, we drop the subindex and just use $(\cdot,\cdot)$ and $\|\cdot\|$, respectively. Thus, if $f, g : \{-1,+1\}^k \to \mathbb{R}$, then

$$(f,g)_\theta = \sum_{x\in\{-1,+1\}^k} f(x)\,g(x)\,v_\theta(x)\,, \quad \|f\|_\theta^2 = \sum_{x\in\{-1,+1\}^k} f^2(x)\,v_\theta(x)\,,$$
$$(f,g) = 2^{-k}\sum_{x\in\{-1,+1\}^k} f(x)\,g(x)\,, \quad \|f\|^2 = 2^{-k}\sum_{x\in\{-1,+1\}^k} f^2(x)\,.$$

We denote the Hilbert space of functions $\{-1,+1\}^k \to \mathbb{R}$ under the inner product $(\cdot,\cdot)$ by J_k.

Fourier transform of clauses. For any $Q \subseteq [k] \equiv \{1,\ldots,k\}$, let $\gamma_Q(x) \stackrel{def}{=} \prod_{i\in Q} x_i$. Under the scalar product defined above (with $\theta = 0$), the functions $\{\gamma_S\}_{S\subseteq[k]}$ form an orthonormal basis for J_k. Moreover, they are exactly the algebraic characters of $\{-1,1\}^k$ with the group operation of pointwise multiplication. Thus, we define the Fourier transform of a function $f \in J_k$, by letting for any $Q \subseteq [k]$,

$$f_Q \stackrel{def}{=} (\gamma_Q, f) = 2^{-k}\sum_{x\in\{-1,+1\}^k} f(x)\gamma_Q(x).$$

Noise operator. Given $\theta \in [-1,1]$, we define the *Bonami - Beckner* operator $\mathrm{T}_\theta : J_k \to J_k$, by

$$(\mathrm{T}_\theta\, f)(x) \stackrel{def}{=} \sum_{y\in\{-1,1\}^k} f(xy)\,v_\theta(y)\,.$$

Notice that $(\mathrm{T}_\theta\, f)(x)$ corresponds to the expected value of $f(\mathbf{x}_\theta)$, where $\mathbf{x}_\theta$ is obtained from x by flipping each coordinate independently with probability $(1-\theta)/2$. Notice that T_1 is just the identity operator and T_0 sends f to the constant function $(f,\gamma_\emptyset)$.

The Bonami-Beckner operator diagonalizes with respect to the Fourier basis, in the sense that $(\mathrm{T}_\theta\,\gamma_Q)(x) = \theta^{|Q|}\gamma_Q(x)$ for any $Q \subseteq [k]$.

More generally, given $h \in [-1,1]^k$, we define $(\mathrm{T}_h\, f)(x) \stackrel{def}{=} \mathbb{E}[f(\mathbf{x}_h)]$, where $\mathbf{x}_h$ is obtained from x by flipping the i^{th} coordinate independently and with probability $\frac{1-h_i}{2}$. Since T_h also diagonalizes with respect to the Fourier basis, one gets $(\mathrm{T}_h\,\gamma_S)(x) = \gamma_S(h)\,\gamma_S(x)\,.$

Discrete derivative and influence. Given a function $f \in J_{k-1}$, we define its *discrete derivative* $f^{(1)} \in J_{k-1}$ as $f^{(1)}(x) = \frac{1}{2}[f(1,x) - f(-1,x)]$. We define analogously $f^{(i)}$ for any other variable index. Finally, the *influence* of the i^{th} variable on f is defined using the norm of the derivative

$$\mathrm{I}_i(f) \stackrel{def}{=} \left\|f^{(i)}\right\|^2.$$

For any $Q \subseteq [k]$, $f_Q^{(i)} = f_{Q\cup\{i\}}$.

3 Main results

3.1 Binary k-CSP ensembles

We assume the following conditions on the ensemble.

1. Permutation symmetry. If φ^π is the Boolean function obtained from φ by permuting its arguments, we require $p(\varphi^\pi) = p(\varphi)$.

2. Balance. The distribution p is supported on Boolean functions such that $\varphi(x_1, \ldots, x_k) = \varphi(-x_1, \ldots, -x_k)$. This condition implies that the odd Fourier coefficients of φ are zero.

3. Feasibility. For each Boolean function φ in the support of p, every partial assignment $(x_1, \ldots, x_{k-1})$ can be extended to a satisfying assignment $(x_0, x_1, \ldots, x_{k-1})$ of φ. This condition implies that $\|\varphi\|^2 \geq 1/2$, and together with the balance condition, implies that all the variables of φ have the same influence, namely, $\mathrm{I}_i(\varphi) = \frac{1-\|\varphi\|^2}{2}$.

4. Dominance of balanced assignments. For every $\theta \in [-1, 1]$,

$$\mathbb{E}_\varphi \log \|\varphi\|_\theta \leq \mathbb{E}_\varphi \log \|\varphi\| .$$

This condition implies that, in a typical random instance, most solutions are balanced in the sense that they have almost as many $+1$'s as -1's.

While our ultimate goal is to exhibit results as $k \to \infty$, the probability distribution p over the functions $\varphi : \{-1, 1\}^k \to \{0, 1\}$ must be defined for *every* k, and some agreement should exist between such probability distributions for different k's. In our work this agreement is given by two conditions concerning the derivative of the clauses in the support of p:

(a) *l_1 norm of the Fourier transform grows at most polynomially in k.* That is, for every $\varphi \in \mathrm{supp}(p)$,

$$\sum_Q \left|\varphi_Q^{(i)}\right| \leq k^a , \tag{1}$$

for some constant a not depending on k.

(b) *'Small weight' Fourier coefficients are small.* There is a constant $C > 0$ (not depending on k) such that for every $\varphi \in \mathrm{supp}\,(p)$,

$$\left\|\mathrm{T}_\theta\, \varphi^{(i)}\right\|^2 \leq e^{-Ck\,(1-\theta)} \left\|\varphi^{(i)}\right\|^2 , \, \theta \in [0, 1] . \tag{2}$$

The above implies in particular, that for any fixed ℓ, there exists $A_\ell > 0$ (independent of k), such that

$$\sum_{1 \leq |Q| \leq \ell} |\varphi_Q|^2 \leq A_\ell e^{-Ck/2} \sum_{|Q| \geq 1} |\varphi_Q|^2. \tag{3}$$

An equivalent formulation of Eq. (2) (with a possibly different constant C) is

$$\left(\mathrm{T}_\theta\, \varphi^{(i)}, \varphi^{(i)}\right) \leq e^{-Ck\,(1-\theta)} \left\|\varphi^{(i)}\right\|^2 , \, \theta \in [0, 1] . \tag{4}$$

Results. An ensemble of binary k-CSP's will be characterized by the following quantities.

$$\frac{1}{\Omega_k} \stackrel{def}{=} \mathbb{E}_\varphi \frac{2\,\mathrm{I}_1(\varphi)}{\|\varphi\|^2} , \qquad \frac{1}{\widehat{\Omega}_k} \stackrel{def}{=} -\mathbb{E}_\varphi \log\left(\|\varphi\|^2\right) .$$

Notice that $\Omega_k \leq \widehat{\Omega}_k$ and $\Omega_k \approx \widehat{\Omega}_k$, whenever the influence is relatively small, or equivalently, when the norm is close to 1.

Proposition 3.1 *A random binary constraint satisfaction instance from the* $\mathrm{CSP}(n, \alpha, p)$ *ensemble is satisfiable, with high probability, if $\alpha < \alpha_{\mathrm{s}}(k)$, where*

$$\Omega_k \, \log 2 \, \{1 + o(1)\} \leq \alpha_{\mathrm{s}}(k) \leq \widehat{\Omega}_k \, \log 2 \, \{1 + o(1)\} .$$

Vice versa, if $\alpha > \alpha_{\mathrm{s}}(k)(1 + o(1))$, then with high probability, a $\mathrm{CSP}(n, \alpha, p)$ *instance is unsatisfiable.*

Given an instance of $\mathrm{CSP}(n,\alpha,p)$, a cluster of solutions is any equivalence class of solutions under the (closure of the) relation $\underline{x} \simeq \underline{x}'$ if $d_{\mathrm{Hamming}}(\underline{x},\underline{x}') \leq d_{\max}$ for some $d_{\max} = o(n)$. The set of solutions is *clustered* if it is partitioned into exponentially many clusters.

Theorem 3.2 *The set of solutions of an instance from the* $\mathrm{CSP}(n,\alpha,p)$ *ensemble is clustered, with high probability, if* $\alpha > \alpha_{\mathrm{d}}(k)$*, where*

$$\alpha_{\mathrm{d}}(k) = \frac{\Omega_k}{k}\left\{\log k + \mathrm{o}(\log k)\right\}.$$

Given a measure $\mu(\underline{x})$ over variable assignments in $\{+1,-1\}^V$, the reconstruction problem is said to be unsolvable if correlations with respect to μ decay rapidly with the distance r on G. More precisely, if $\mu_{i,\sim r}$ denotes the joint distribution of x_i and $\{x_j : d_G(i,j) \geq r\}$, then $\lim_{r\to\infty}\limsup_{n\to\infty} \mathbf{E}\|\mu_{i,\sim r} - \mu_i\mu_{\sim r}\|_{\mathrm{TV}} = 0$.

Theorem 3.3 *Let* $\mu(\underline{x})$ *be the uniform measure over solutions of an instance from the* $\mathrm{CSP}(n,\alpha,p)$ *ensemble. The reconstruction problem is solvable for* μ *if* $\alpha > \alpha_{\mathrm{r}}(k)$*, where*

$$\alpha_{\mathrm{r}}(k) = \frac{\Omega_k}{k}\left\{\log k + \mathrm{o}(\log k)\right\}.$$

Vice versa, the reconstruction problem is unsolvable if $\alpha < \alpha_{\mathrm{r}}(k)$.

Thus, a key result of the present paper is that $\alpha_{\mathrm{d}}(k)$ and $\alpha_{\mathrm{r}}(k)$ do *coincide for a large family of ensembles* (up to the slackness, in the second order terms, of our bounds).

Example: 2-coloring hypergraphs. Let us consider the ensemble of CSP's consisting of clauses of the type φ, where $\varphi(x_1,\dots,x_k) = \mathbb{I}\left(\sum x_i \notin \{-k,k\}\right)$. The $\mathrm{CSP}(n,\alpha,p)$ in this case, corresponds to the distribution of 2-colorings of a random hypergraph on n vertices and αn edges, with edge size k, and each edge chosen independently and uniformly at random.

The conditions 1-3 clearly hold for this model and the dominance of balance assignments follows after checking that $\|\varphi\|_\theta = 1 - \left(\frac{1+\theta}{2}\right)^k - \left(\frac{1-\theta}{2}\right)^k$ maximizes at $\theta = 0$. To establish the conditions (1), notice that $\varphi^{(i)}_Q = -\frac{1}{2^k}[1-(-1)^{|Q|}]$, which clearly implies that the l_1 norm of the fourier transform is bounded. To check (2), notice that $\frac{\left(\mathrm{T}_\theta\,\varphi^{(i)},\varphi^{(i)}\right)}{\|\varphi^{(i)}\|^2} = \left(\frac{1+\theta}{2}\right)^{k-1} - \left(\frac{1-\theta}{2}\right)^{k-1} \leq e^{-k(1-\theta)/2}$ for all $\theta \in [0,1]$.

An easy computation shows that $\Omega_k = 2^{k-1}-1$ and $\frac{1}{\widehat{\Omega}_k} = -\log(1-2^{-k+1})$, therefore we have:

	Reconstruction - Clustering	Lower bound satisfiability	Upper bound satisfiability
2-coloring	$\frac{2^{k-1}}{k}\,[\log k + \mathrm{o}(\log k)]$	$2^{k-1}\log 2\,[1+\mathrm{o}(1)]$	$2^{k-1}\log 2\,[1+\mathrm{o}(1)]$

Example: Not All Equal $k-$SAT. Let us consider now an ensemble of CSP's consisting of clauses of type $\{\varphi_s\}_{s\in\{+1,-1\}^k}$, where $\varphi_s(x_1,\dots,x_k) = \mathbb{I}\left(\sum x_i s_i \notin \{-k,k\}\right)$ and $p(\varphi_s) = 2^{-k}$ for each $s \in \{+1,-1\}^k$. In this case, the CSP (n,α,p) model corresponds to the distribution of NAE $k-$SAT instances for a random formula in n variables, consisting of αn random clauses, each with k literals.

For this model, the conditions 1-3 are easily verified. The dominance of balance assignments follows from

$$\mathbb{E}_s \log\|\varphi\|_\theta \leq \log \mathbb{E}_s\|\varphi\|_\theta = \log \mathbb{E}_s\left(1 - \prod_{i=1}^k \frac{1+s_i\theta}{2} - \prod_{i=1}^k \frac{1-s_i\theta}{2}\right) = \mathbb{E}_s \log\|\varphi\|.$$

On the other hand, the Fourier expansion of φ_s is given by $\varphi_{s,Q} = -2^{-k}[\gamma_Q(s)+\gamma_Q(-s)]$. In particular $|\varphi_{s,Q}|^2 = 2^{-k}[1+(-1)^{|Q|}]$, so that both Eqs. (1) and (2) hold along the same lines as the previous example. Indeed, in this case we get the same values for Ω_k and $\widehat{\Omega}_k$, so that, we have:

	Reconstruction - Clustering	Lower bound satisfiability	Upper bound satisfiability
NAE-SAT	$\frac{2^{k-1}}{k}\,[\log k + \mathrm{o}(\log k)]$	$2^{k-1}\log 2\,[1+\mathrm{o}(1)]$	$2^{k-1}\log 2\,[1+\mathrm{o}(1)]$

Example: k-XOR formulas. For an even integer k, the k-XOR ensemble (k even) consists of clauses of type $\{\varphi_\epsilon\}_{\epsilon=1,-1}$, where $\varphi_\epsilon(x_1,\dots,x_k) = \frac{1}{2}\left(\gamma_\emptyset + \epsilon\gamma_{[k]}\right)$. In this case, the CSP (n,α,p) model corresponds to a system of αn random linear equations in $\mathbb{Z}_2$, in which every equation involves k randomly chosen variables (with replacement) from a total of n possible variables.

Conditions 1-3 hold for k even, and the dominance of balanced assignments condition follows from the fact that $\mathbb{E}_\varphi \log\|\varphi\|_\theta = \frac{1}{2}\log\left(\frac{1-\theta^{2k}}{4}\right)$, which is clearly maximized at $\theta = 0$. The condition on Fourier expansion of clauses for this model is straightforward: The Fourier expansion of φ_ϵ is concentrated at $\emptyset$ and $[k]$, so that the Eq. (1) holds with $a = 0$ and the Eq. (1) holds with $C = 1$.

In this case, we have that $\Omega_k = 1$, while $\widehat{\Omega}_k = 1/\log 2$. Therefore, we have:

	Reconstruction - Clustering	Lower bound satisfiability	Upper bound satisfiability
XOR-SAT	$\frac{1}{k}\left[\log k + \mathrm{o}(\log k)\right]$	$\log 2 + \mathrm{o}(1)$	$1 + \mathrm{o}(1)$

We remark here that, in the case of XOR-SAT, the clustering and satisfiability thresholds can be determined *exactly* by exploiting the underlying group structure [MRZ03, CD+03] (see [MM09] for a discussion of the reconstruction problem in XOR-SAT).

3.2 q-ary ensembles: graph coloring

The following result concerning the colorability and clustering of proper colorings were proved by Achlioptas and Naor [AN05] and Achlioptas and Coja-Oghlan [AC08].

Theorem 3.4 *(Graph q-colorability [AN05]) A random graph with n vertices and $n\alpha$ edges is satisfiable with high probability if $\alpha < \alpha_{\rm s}(q)$, where*

$$\alpha_{\rm s}(q) = q\left[\log q + o_q(1)\right] .$$

Vice versa, if $\alpha > \alpha_{\rm s}(q)(1 + o_q(1))$, such a graph is with high probability uncolorable.

Theorem 3.5 *(Clustering of q-colorings [AC08]) The set of proper q-colorings of random graph with n vertices and $n\alpha$ edges is clustered with high probability if $\alpha > \alpha_{\rm d}(q)$, where*

$$\alpha_{\rm d}(q) = \frac{q}{2}\left[\log q + o(\log q)\right].$$

One of our main results is to prove a corresponding reconstruction theorem for this model as follows.

Theorem 3.6 *(Graph q-coloring reconstruction) Let $\mu(\underline{x})$ be the uniform measure over of proper q-colorings of random graph with n vertices and $n\alpha$ edges. For q large enough, the reconstruction problem is solvable for μ if $\alpha > \alpha_{\rm r}(q)$, where*

$$\alpha_{\rm r}(k) = \frac{q}{2}\left[\log q + \log\log q + O\left(1\right)\right] .$$

Vice versa, the reconstruction problem is unsolvable, with high probability, if $\alpha < \alpha_{\rm r}(q)$.

3.3 General strategy

The results described in the previous section are of three types: bounds on the satisfiability thresholds, cf. Proposition 3.1 and Theorem 3.4; on the clustering threshold, cf. Theorems 3.2 and 3.5; on the reconstruction threshold, cf. Theorems 3.3 and 3.6. The proof strategy is as follows.

The satisfiability threshold can be upper bounded using the first moment of the number of solutions, and lower bounded using the second moment method. This technique is by now discussed in detail in [AM02, AN05, ANP05]; we describe its application to the general CSP(n,α,p) ensemble is done in Appendix A.

The clustering threshold can be upper bounded through an analysis of the recursive 'whitening' process that associates to each cluster a single configuration in an extended space [AR06]. The improved bounds in Theorems 3.2 and 3.5 can be obtained by approximating the CSP ensemble with an appropriate 'planted' ensemble [AC08]. Since this approach is explained in detail in [AC08], we will only present the various technical steps.

The reconstruction threshold is characterized via a three-step procedure:

(1) Bound the reconstruction threshold for an appropriate ensemble of (infinite) tree instances, i.e. CSP instances for which the associated factor graph is an infinite Galton-Watson tree. In the case of proper q-colorings, a sharp characterization was obtained independently by two groups in the past year [BVV07, Sly08]. In Section 4 we prove sharp bounds on tree reconstruction for binary CSPs. The proof amounts to deriving an exact distributional recursion for the so-called belief process, and carefully bounding its asymptotic behavior.
(2) Given two 'balanced' solutions $\underline{x}^{(1)}$, $\underline{x}^{(2)}$ (a solution is balanced if each possible variable value is taken on the same number of vertices), define their *joint type* $\nu(x, y)$ as the matrix such that the fraction of vertices i with $x_i^{(1)} = x$ and $x_i^{(2)} = y$ is equal to $\nu(x, y)$. Consider the number $Z_{\rm b}(\nu)$ of balanced solution pairs $\underline{x}_1$, $\underline{x}_2$ with joint type ν. One has to show that $\mathbb{E}\, Z_{\rm b}(\nu)$ is exponentially dominated by its value at the uniform type $\overline{\nu}(x, y) = 1/q^2$ (with $q = 2$ for binary CSPs). More precisely $\mathbb{E}\, Z_{\rm b}(\nu) \doteq \exp\{n\Phi(\nu)\}$ with Φ achieving its unique maximum at $\overline{\nu}$.

This is also a crucial step in the second moment method. It was accomplished in [AN05] for proper q-colorings of random graphs. In the case of binary CSPs, we prove this estimate in Section A.
(3) Prove that the above imply that the set of solutions of a random instance is, with high probability, *roughly spherical*. By this we mean that the joint type ν_{12} of two uniformly random solutions $\underline{x}^{(1)}, \underline{x}^{(2)}$ satisfies $||\nu_{12} - \overline{\nu}||_{\rm TV} \le \delta$ with high probability for all $\delta > 0$. Notice that this implication requires bounding the expected ratio of $Z_{\rm b}(\nu)$ to the total number of solution pairs. We prove that the implication nevertheless holds in Section 5 for q-colorings. The argument for binary CSP's is completely analogous, and we omit it.

Finally, it was proved in [GM07] that, under such a sphericity condition, graph reconstruction and tree reconstruction are equivalent, which finishes the proof of Theorems 3.3 and 3.6.

Notice that the techniques used for the clustering and reconstruction thresholds are very different. Thus it is a surprising (and arguably deep) phenomenon that they do coincide as far as the present techniques can tell.

4 Tree ensembles and tree reconstruction for binary k-CSP ensembles

In this section we define tree ensembles and prove estimates about the corresponding tree reconstruction thresholds.

4.1 The tCSP(α, p) ensemble

The ensemble tCSP(α, p) is defined by $\alpha \in \mathbb{R}_+$ and a distribution p over Boolean functions $\varphi : \{-1, +1\}^k \to \{0, 1\}$. We assume the conditions on the distribution p introduced in Section 3.1. An (infinite) instance from this ensemble is generated starting by a root variable node $\o$, drawing an integer $\eta \stackrel{\mathcal{D}}{=} \text{Poisson}(k\alpha)$ and connecting $\o$ to η function nodes $\{1, \dots, \eta\}$. Each function node has degree k, and each of its $k - 1$ descendants is the root of an independent infinite tree. Finally, each function node a is associated independently, with a random clause φ drawn according to p.

A uniform solution for such an instance is sampled by drawing the root value $\mathbf{x}_{\o} \in \{-1, +1\}$ uniformly at random. The values of descendants of each variable node i are then drawn recursively. If the function node a connects i to $i_1, \dots, i_{k-1}$, then the values $\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_k}$ are sampled uniformly from those that satisfy the clause in a, that is, such that the quantity $\varphi\left(x_i, x_{i_1}, \dots, x_{i_{k-1}}\right)$ is equal to 1.

By the *balance* condition, this procedure can be shown to be equivalent to sampling a solution according to the 'free boundary Gibbs measure.' The latter is a distribution over solutions of the entire (infinite) tCSP formula defined by considering the unifom distribution over solutions of the first ℓ generations of the tree, and then letting $\ell \to \infty$.

4.2 Reconstruction

Given any fixed tree ensemble T, let $\mathbf{x}$ be a random satisfying assignment for T according to the distribution described previously. We denote by $\mathbf{x}_\ell$ the value of $\mathbf{x}$ at the variables at generation ℓ, and in the case that the root degree is 1, we denote by $\mathbf{x}_{0,1}, \dots, \mathbf{x}_{0,k-1}$, the value at the variable nodes connected to the unique child of the root. Also, we use η_0 for the root degree of T. If the tree ensemble T has root degree $\eta_0 = d$, we denote by T_i, $i = 1, \dots, d$, the subtree generated by the root, its i^{th} children and its descendents. If $\eta_0 = 1$, we denote by T_i', $i = 1, \dots, k-1$, the subtree generated by the i^{th} child of the root's child and its descendents.

Finally, because the tree ensemble T could be random (for instance we denote by $\mathbf{T}$ a random tCSP$\,(\alpha, p)$), we will use $\boldsymbol{E}$ for expectation respect to $\mathbf{T}$, and $\langle\cdot\rangle_T$ for expectation respect to $\mathbf{x}$ (given $\mathbf{T}$) and $\mathbb{E}$ for expectation respect to any other independent random variable (adding, if not in context, a subindex to indicate such random variable).

Reconstruction: For a fixed tree ensemble T, let $\mu_{\emptyset,\ell}$ be the joint distribution of $(\mathbf{x}_0, \mathbf{x}_\ell)$ and let $\mu_\emptyset$, μ_ℓ be the marginal distribution of $\mathbf{x}_0$ and $\mathbf{x}_\ell$ respectively. The reconstruction rate for T is defined as the quantity $\left\|\mu_{\emptyset,\ell}(\cdot,\cdot) - \mu_\emptyset(\cdot)\,\mu_\ell(\cdot)\right\|_{\mathrm{TV}}$. We say that the reconstruction problem for T is *tree-solvable* if

$$\liminf_{\ell\to\infty} \left\|\mu_{\emptyset,\ell}(\cdot,\cdot) - \mu_\emptyset(\cdot)\,\mu_\ell(\cdot)\right\|_{\mathrm{TV}} > 0.$$

Analogously, if $\mathbf{T}$ is a random tCSP(α, p), we define the reconstruction rate of $\mathbf{T}$ as $\mathbf{E}\left\|\mu_{\emptyset,\ell}(\cdot,\cdot) - \mu_\emptyset(\cdot)\,\mu_\ell(\cdot)\right\|_{\mathrm{TV}}$, and we say that the reconstruction problem for $\mathbf{T}$ is *tree-solvable*

$$\liminf_{\ell\to\infty} \mathbf{E}\left\|\mu_{\emptyset,\ell}(\cdot,\cdot) - \mu_\emptyset(\cdot)\,\mu_\ell(\cdot)\right\|_{\mathrm{TV}} > 0.$$

Bias, compatibility: Given a satisfying assignment x_ℓ for the variables at generation ℓ, define the 'bias' of the root, restricted to the value of the variables at level ℓ, as

$$h_T(x_\ell) \stackrel{def}{=} \langle \mathbf{x}_0 \,| \mathbf{x}_\ell = x_\ell \rangle_T .$$

Throughout the next proofs we will study $h_T(x_\ell)$, for x_l random and subject to different kind of distributions. Notice that under the balance condition $\left\|\mu_{\emptyset,\ell}(\cdot,\cdot) - \mu_\emptyset(\cdot)\,\mu_\ell(\cdot)\right\|_{\mathrm{TV}} = \langle |h_T(\mathbf{x}_\ell)| \rangle_T$.

Now, let $D_T(x_\ell) \stackrel{def}{=} \{x\}$ if $h_T(x_\ell) = x$, $D_T(x_\ell) \stackrel{def}{=} \{-1, 1\}$ if $|h_T(x_\ell)| < 1$. Observe that $D_T(x_\ell)$ consists of the values of the root that are compatible with the assignment x_ℓ for the variables at generation l.

Domain of clauses: Given a binary function $\varphi(x_0, \ldots, x_{k-1})$, define the partial solution sets

$$S^+(\varphi) \stackrel{def}{=} \{(x_1, , x_{k-1}) : \varphi(1, x_1, \ldots, x_{k-1}) = 1\},$$
$$S^-(\varphi) \stackrel{def}{=} \{(x_1, , x_{k-1}) : \varphi(-1, x_1, \ldots, x_{k-1}) = 1\},$$

$$\Lambda^+(\varphi) \stackrel{def}{=} S^+(\varphi)\setminus S^-(\varphi), \qquad \Lambda^-(\varphi) \stackrel{def}{=} S^-(\varphi)\setminus S^+(\varphi)$$

If the clause φ is balanced and feasible, we have that $|S^+(\varphi)| = |S^-(\varphi)| = 2^{k-1}\|\varphi\|^2$ and $|\Lambda^+(\varphi)| = |\Lambda^-(\varphi)| = 2^k\,\mathrm{I}_1(\varphi)$.

Theorem 4.1 *The reconstruction problem for the ensemble tCSP(α, p) is* tree-solvable *if and only if* $\alpha > \alpha_{\mathrm{tree}}(k)$ *where*

$$\alpha_{\mathrm{tree}}(k) = \frac{\Omega_k}{k}\{\log k + o(\log k)\}.$$

Proof. *Upper bound:*

Given a tree ensemble T, the rate of 'naive reconstruction' for T is defined as

$$z_\ell(T) \stackrel{def}{=} \langle \mathbb{I}[h_T(\mathbf{x}_\ell) = 1] \rangle_T \;\; (= \langle \mathbb{I}[h_T(\mathbf{x}_\ell) = -1] \rangle_T \text{ by the balance condition}),$$

which indicates the probability that a random assignment for the variables at generation ℓ, distributed as $\mathbf{x}_\ell$, fixes the root to be equal to 1 (or -1). It is easy to see that $\langle |h_T(\mathbf{x}_\ell)| \rangle_T \geq z_\ell(T)$. Observe also, that for any $x, y \in \{-1, 1\}$,

$$\langle \mathbb{I}[h_T(\mathbf{x}_\ell) = x] \,| \mathbf{x}_0 = y \rangle_T = 2 z_\ell(T)\,\delta_{x,y}. \tag{5}$$

Thus, our objective is to show that in an appropiate regime of the parameter α, the quantity $\mathbf{E}[z_\ell(\mathbf{T})]$ remains bounded away from zero as $\ell \to \infty$, implying tree-solvability of the reconstruction problem in such regime. Indeed, this implies tree-solvability by 'naive reconstruction', i.e. by the procedure that assigns to the root any value compatible with the values at generation ℓ. By notational convenience, define

$$z_\ell(\alpha) = 2\mathbf{E}[z_\ell(\mathbf{T})] \text{ and } \widehat{z}_\ell(\alpha) = 2\mathbf{E}[z_\ell(\mathbf{T}) \,|\eta_0 = 1].$$

Now, notice that for a tree ensemble T with root degree $\eta_0 = d$, and any assignment x_ℓ for the variables at generation ℓ, $h_T(x_\ell) = 1$ iff $h_T(x_\ell \upharpoonright T_i) = 1$ for some $i = 1, \ldots, d$, so that

$$\begin{aligned} 2z_\ell(T) &= \left\langle 1 - \prod_{i=1}^{d} \left(1 - \mathbb{I}\left[h_{T_i}(\mathbf{x}_\ell \upharpoonright T_i) = 1\right]\right) | \mathbf{x}_0 = 1 \right\rangle_T \\ &= 1 - \prod_{i=1}^{d} \left\langle \left(1 - \mathbb{I}\left[h_{T_i}(\mathbf{x}_\ell) = 1\right]\right) \middle| \mathbf{x}_0 = 1 \right\rangle_{T_i} \quad \text{(By the tree Markov property)} \\ &= 1 - \prod_{i=1}^{d} \left(1 - 2z_\ell(T_i)\right) . \end{aligned}$$

Therefore, averaging over T, we get

$$\begin{aligned} z_\ell(\alpha) &= \mathbb{E}_\eta \left[1 - \prod_{i=1}^{\eta} \left(1 - \widehat{z}_\ell(\alpha)\right)\right], \, \eta \sim \text{Poisson}(k\alpha) \\ &= 1 - \exp\left(-k\alpha \widehat{z}_\ell(\alpha)\right). \end{aligned}$$

On the other hand, given a tree ensemble T with root degree $\eta_0 = 1$ and with the clause φ assigned to the root's child, we have that for any satisfying assignment x_ℓ for the variables at generation ℓ, $h_T(x_\ell) = 1$ iff

$$\prod_{i=1}^{k-1} D_{T_i'}\left(x_{\ell-1}^{(i)}\right) \subseteq \Lambda^+(\varphi), \tag{6}$$

where $x_{\ell-1}^{(i)}$ is the assignment $x_\ell \upharpoonright T_i'$ for the variables at generation $\ell - 1$ in the subtree T_i'. Observe that (6) holds, in particular, if for some $a = (a_1, \ldots, a_{k-1}) \in \Lambda^+(\varphi)$, $h_{T_i'}\left(x_{\ell-1}^{(i)}\right) = a_i$ for $i = 1, \ldots, k-1$. Therefore, if $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_{k-1})$ denotes a random uniform vector from $S^+(\varphi)$, we have

$$\begin{aligned} z_\ell(T) &\geq \frac{1}{2} \sum_{a \in \Lambda^+(\varphi)} \left\langle \prod_{i=1}^{k-1} \mathbb{I}\left[h_{T_i'}\left(\mathbf{x}_{\ell-1}^{(i)}\right) = a_i\right] | \mathbf{x}_0 = 1 \right\rangle_T \\ &= \frac{1}{2} \sum_{a \in \Lambda^+(\varphi)} \mathbb{E}_\mathbf{y} \prod_{i=1}^{k-1} \left\langle \mathbb{I}\left[h_{T_i'}(\mathbf{x}_{\ell-1}) = a_i\right] | \mathbf{x}_0 = y_i \right\rangle_{T_i'} \quad \text{(By the tree Markov property)} \\ &= \frac{|\Lambda^+(\varphi)|}{|S^+(\varphi)|} \prod_{i=1}^{k-1} 2z_{\ell-1}(T_i') \quad \text{(By Eq. (5))}, \end{aligned}$$

which implies, after averaging over T, that

$$\widehat{z}_\ell(\alpha) \geq \mathbb{E}_\varphi \left[\frac{2\,\mathrm{I}_1(\varphi)}{\|\varphi\|^2}\right] \left(z_{\ell-1}(\alpha)\right)^{k-1} = \frac{\left(z_{\ell-1}(\alpha)\right)^{k-1}}{\Omega_k},$$

which leads to the recursion $z_\ell(\alpha) \geq 1 - \exp\left(-k\alpha \left(z_{\ell-1}(\alpha)\right)^{k-1} / \Omega_k\right)$. Now, it is standard to verify that this recursion implies that $z_\ell(\alpha)$ is, for all ℓ, greater or equal than the maximum of the fixed points of the function $g(z) = 1 - \exp\left(-k\alpha z^{k-1}/\Omega_k\right)$ in the interval $[0, 1]$. The minimum value of α for which such fixed point is positive is given by

$$\alpha^* = \frac{\Omega_k \left(1 + u\left(1 + \frac{1}{u}\right)^{k-2}\right)}{k(k-1)},$$

where u is the unique solution of the equation $u = (k-1)\log(1+u)$. In particular, asymptotically in k, we have that $\alpha^* = \frac{\Omega_k}{k}(\log k + \mathrm{o}(\log k))$, which implies the upper bound for α_{tree}.

Lower bound:

The matching lower bound on $\alpha_{\rm tree}(k)$ requires a more elaborate proof; we first prove three lemmas, before returning to complete the lower bound proof. $\square$

Given a tree ensemble T, let $\mathbf{x}_\ell^+ \stackrel{\mathcal{D}}{=} (\mathbf{x}_\ell \,|\mathbf{x}_0 = 1)$ and $\mathbf{x}_\ell^- \stackrel{\mathcal{D}}{=} (\mathbf{x}_\ell \,|\mathbf{x}_0 = -1)$. When the tree ensemble is not clear in the definition of $\mathbf{x}_\ell^+$ (or $\mathbf{x}_\ell^-$), we add a subindex indicating the tree ensemble from where it is defined. Notice that, if μ^+ and μ^- are the distributions of $\mathbf{x}_\ell^+$ and $\mathbf{x}_\ell^-$ respectively, then

$$\frac{d\mu^-}{d\mu^+} = \frac{1-h_T(x_l)}{1+h_T(x_l)}. \tag{7}$$

By the balance condition, it's clear that

$$h_T\left(\mathbf{x}_\ell^+\right) \stackrel{\mathcal{D}}{=} -h_T\left(\mathbf{x}_\ell^-\right). \tag{8}$$

Also, it is easy to show that $\left\langle h_T\left(\mathbf{x}_\ell^+\right)\right\rangle_T = \left\langle \left[h_T\left(\mathbf{x}_\ell\right)\right]^2\right\rangle_T$ (and therefore $\left[R_l\left(T\right)\right]^2 \leq \left\langle h_T\left(\mathbf{x}_\ell^+\right)\right\rangle_T \leq R_l\left(T\right)$), so that non-reconstructibility for T is equivalent to the condition $\lim_{\ell\to\infty} \left\langle h_T\left(\mathbf{x}_\ell^+\right)\right\rangle_T = 0$ (see [MP03]). Similarly, if $\mathbf{T}$ is a random tCSP (α, p) ensemble, non-reconstructibility for $\mathbf{T}$, is equivalent to the condition $\lim_{\ell\to\infty} \mathbf{E}\left[\left\langle h_{\mathbf{T}}\left(\mathbf{x}_\ell^+\right)\right\rangle_{\mathbf{T}}\right] = 0$.

Lemma 4.2 ***(a)*** *Given a tree ensemble T with root degree $\eta_0 = d$, we have*

$$\left[\frac{1-h_T\left(\mathbf{x}_\ell^+\right)}{1+h_T\left(\mathbf{x}_\ell^+\right)}\right] \stackrel{\mathcal{D}}{=} \prod_{i=1}^{d}\left[\frac{1-h_{l,i}}{1+h_{l,i}}\right], \tag{9}$$

where $(h_{l,i})_{i=1}^d$ are independent random variables such that $h_{l,i} \stackrel{\mathcal{D}}{=} h_{T_i}\left(\mathbf{x}_\ell^+\right)$.

(b) *Given a tree ensemble T with root degree $\eta_0 = 1$ and with the clause φ assigned to the unique child of the root, we have that*

$$\left[\frac{1-h_T\left(\mathbf{x}_{\ell+1}^+\right)}{1+h_T\left(\mathbf{x}_{\ell+1}^+\right)}\right] \stackrel{\mathcal{D}}{=} \frac{\mathrm{T}_{h_l}\,\varphi(-1,\mathbf{s})}{\mathrm{T}_{h_l}\,\varphi(1,\mathbf{s})}, \tag{10}$$

where $\mathbf{s} \sim \mathrm{Unif}\left(S^+\left(\varphi\right)\right)$ and $h_l = (h_{l,i})_{i=1}^{k-1}$ are independent random variables such that $h_{l,i} \stackrel{\mathcal{D}}{=} h_{T_i'}\left(\mathbf{x}_l^+\right)$.

Proof. This recursion follows straightforwardly from the recursive definition of tree formulae. The balance condition on clauses implies

$$\frac{1-h_T\left(\mathbf{x}_l^+\right)}{1+h_T\left(\mathbf{x}_l^+\right)} = \frac{\left\langle \mathbb{I}\left[\mathbf{x}_l = \mathbf{x}_l^+\right]|\mathbf{x}_0 = -1\right\rangle_T}{\left\langle \mathbb{I}\left[\mathbf{x}_l = \mathbf{x}_l^+\right]|\mathbf{x}_0 = 1\right\rangle_T}.$$

Therefore, if the root degree of T is $\eta_0 = d$, we have by the tree Markov property that

$$\frac{1-h_T\left(\mathbf{x}_l^+\right)}{1+h_T\left(\mathbf{x}_l^+\right)} = \prod_{i=1}^{d}\frac{\left\langle \mathbb{I}\left[\mathbf{x}_l = \mathbf{x}_l^+ \restriction T_i\right]|\mathbf{x}_0 = -1\right\rangle_{T_i}}{\left\langle \mathbb{I}\left[\mathbf{x}_l = \mathbf{x}_l^+ \restriction T_i\right]|\mathbf{x}_0 = 1\right\rangle_{T_i}},$$

and the last expression has the same distribution as $\prod_{i=1}^{d}\frac{1-u_{l,i}}{1+u_{l,i}}$, due to the fact that $\left(\mathbf{x}_l^+ \restriction T_i\right)_{i=1}^d$ are independent random assignments for the variables at generation l of T_i, such that $\mathbf{x}_l^+ \restriction T_i \stackrel{\mathcal{D}}{=} \mathbf{x}_{l,T_i}^+$. This proves Eq. (9). Now, if the root degree of T is $\eta_0 = 1$, define $\left(\widetilde{\mathbf{x}}_{l,i}^+\right)_{i=1}^{k-1}$ to be independent random assignments for the variables at generation l of the subtrees T_i', such that $\widetilde{\mathbf{x}}_{l,i}^+ \stackrel{\mathcal{D}}{=} \mathbf{x}_{l,T_i'}^+$. By the tree Markov property, we have that $\left(\mathbf{x}_{l+1}^+ \restriction T_i'\right)_{i=1}^{k-1} \stackrel{\mathcal{D}}{=} \left(\mathbf{s}_i\widetilde{\mathbf{x}}_{l,i}^+\right)_{i=1}^{k-1}$

where $\mathbf{s} \sim \text{Unif}\, S^+(\varphi)$. Using once more the tree Markov property, we get

$$\left[\frac{1-h_T\left(\mathbf{x}_{\ell+1}^+\right)}{1+h_T\left(\mathbf{x}_{\ell+1}^+\right)}\right] = \frac{\sum_y \varphi(-1,y) \prod_{i=1}^{k-1} \left\langle \mathbb{I}\left[\mathbf{x}_l = \mathbf{s}_i \widetilde{\mathbf{x}}_{l,i}^+\right] |\mathbf{x}_0 = y_i \right\rangle_{T_i'}}{\sum_y \varphi(-1,y) \prod_{i=1}^{k-1} \left\langle \mathbb{I}\left[\mathbf{x}_l = \mathbf{s}_i \widetilde{\mathbf{x}}_{l,i}^+\right] |\mathbf{x}_0 = y_i \right\rangle_{T_i'}}$$
$$= \frac{\mathrm{T}_{h_l}\, \varphi(-1,\mathbf{s})}{\mathrm{T}_{h_l}\, \varphi(1,\mathbf{s})},$$

which is precisely Eq. (10). □

The first step of the above recursion can be analyzed exactly.

Lemma 4.3 *If $\mathbf{T}$ is a random* tCSP(α, p) *ensemble, then the random variable $h_{\mathbf{T}}\left(\mathbf{x}_1^+\right)$ takes values in $\{0,1\}$ and, if $\alpha < (1-\delta)(\Omega_k \log k)/k$, we have $\mathbf{E}\, h_{\mathbf{T}}\left(\mathbf{x}_1^+\right) \leq 1 - k^{-1+\delta}$.*

Proof. If T is a tree ensemble with root degree $\eta_0 = 1$ and clause φ assigned to the root's child, from the part b of lemma 4.2, we have that $\frac{1-h_T\left(\mathbf{x}_1^+\right)}{1+h_T\left(\mathbf{x}_1^+\right)} \overset{\mathcal{D}}{=} \varphi(-1,\mathbf{s})$ where $\mathbf{s} \sim \text{Unif}\left(S^+(\varphi)\right)$ (notice that $h_{0,i} \equiv 1$). Therefore, it follows that $h_T\left(\mathbf{x}_1^+\right) = 1$ w.p. $\frac{\left|\Lambda^+(\varphi)\right|}{\left|S^+(\varphi)\right|} = 1/\Omega_k$ and $h_T\left(\mathbf{x}_1^+\right) = 0$ otherwise. Therefore, if T is a tree ensemble with root degree $\eta_0 = d$, it follows from the part a of lemma 4.2 that $h_T\left(\mathbf{x}_1^+\right) = 1$ w.p. $1-(1-1/\Omega_k)^d$ and $h_T\left(\mathbf{x}_1^+\right) = 0$ otherwise. This implies then that $h_{\mathbf{T}}\left(\mathbf{x}_1^+\right)$ is supported at $\{0,1\}$ and $\mathbf{E} h_{\mathbf{T}}\left(\mathbf{x}_1^+\right) = 1 - \exp\left(-k\alpha\left(1-1/\Omega_k\right)\right)$. The conclusion follows straightforwardly.

□

For subsequent steps we track the averages, $h_\ell^{\text{ave}} \overset{def}{=} \mathbf{E}\, \left\langle h_{\mathbf{T}}\left(\mathbf{x}_l^+\right)\right\rangle_{\mathbf{T}}$ and $\widehat{h}_\ell^{\text{ave}} \overset{def}{=} \mathbf{E}\, \left[\left\langle h_{\mathbf{T}}\left(\mathbf{x}_l^+\right)\right\rangle_{\mathbf{T}} |\eta_0 = 1\right]$, using the following bounds.

Lemma 4.4 *For any $\ell \geq 0$ we have*

$$h_\ell^{\text{ave}} \leq 1 - e^{-2k\alpha \widehat{h}_\ell^{\text{ave}}}, \qquad \widehat{h}_{\ell+1}^{\text{ave}} \leq \frac{1}{2} F_k(h_\ell^{\text{ave}}) + \frac{1}{2} R_k(\sqrt{h_\ell^{\text{ave}}}), \tag{11}$$

$$F_k(\theta) \overset{def}{=} 2\mathbb{E}_\varphi\left[\frac{(\varphi^{(1)}, \mathrm{T}_\theta\, \varphi^{(1)})}{\|\varphi\|^2}\right], \qquad R_k(\theta) \overset{def}{=} 2\mathbb{E}_{\varphi\mathbf{i}}\left[\frac{2\,\mathrm{I}_1(\varphi)}{\|\varphi\|^2} \sum_{Q\subseteq[k-1]} |\left(\varphi^{(1)}, \gamma_Q\right)|\, \theta^{\max(|Q|,2)}\right], \tag{12}$$

Finally, if h_ℓ is supported on non-negative values, then

$$\widehat{h}_\ell^{\text{ave}} \leq F_k(h_\ell^{\text{ave}}). \tag{13}$$

Proof. We will say that a random variable $\mathbf{X} \in [-1,+1]$ is 'consistent,' if $\mathbb{E}\, f(-\mathbf{X}) = \mathbb{E}\left[\left(\frac{1-\mathbf{X}}{1+\mathbf{X}}\right)\, f(\mathbf{X})\right]$for every function f such that the expectation values exist. A useful preliminary remark [MM06] is that the random variable $h_T\left(\mathbf{x}_l^+\right)$ is consistent (no matter the tree ensemble). In fact, this follows directly from the Eqs. (7) and (8) above. A number of properties of consistent random variables can be found in [RU08]. Let us now consider the first inequality. If T is a tree ensemble with root degree $\eta_0 = d$, it is immediate to from Eq. (9), that

$$\left\langle \left(\frac{1-h_T\left(\mathbf{x}_l^+\right)}{1+h_T\left(\mathbf{x}_l^+\right)}\right)^{1/2}\right\rangle_T = \prod_{i=1}^d \left\langle \left(\frac{1-h_{T_i}\left(\mathbf{x}_l^+\right)}{1+h_{T_i}\left(\mathbf{x}_l^+\right)}\right)^{1/2}\right\rangle_{T_i}.$$

It is possible to show that consistency implies $E\,X = E\,X^2$ and $E\left(\frac{1-X}{1+X}\right)^{1/2} = E\sqrt{1-X^2}$ (through the test functions $f(x) = x(1+x)$ and $f(x) = x(1+x)^{1/2}(1-x)^{-1/2}$), we thus have

$$\sqrt{1-\left\langle h_T\left(\mathbf{x}_l^+\right)\right\rangle_T} \geq \left\langle \sqrt{1-\left[h_T\left(\mathbf{x}_l^+\right)\right]^2}\right\rangle_T = \prod_{i=1}^d \left\langle \sqrt{1-\left[h_{T_i}\left(\mathbf{x}_l^+\right)\right]^2}\right\rangle_{T_i} \geq \prod_{i=1}^d \left(1-\left\langle h_{T_i}\left(\mathbf{x}_l^+\right)\right\rangle_{T_i}\right).$$

This implies in particular, if $\mathbf{T}$ is a random tCSP (α, p),

$$\sqrt{1-\mathbf{E}\left\langle h_{\mathbf{T}}\left(\mathbf{x}_l^+\right)\right\rangle_{\mathbf{T}}} \geq \mathbb{E}_\eta\left[\prod_{i=1}^{\eta}\left(1-\mathbf{E}\left[\left\langle h_{\mathbf{T}}\left(\mathbf{x}_l^+\right)\right\rangle_{\mathbf{T}} | \eta_0=1\right]\right)\right], \eta \sim \text{Poisson}\left(k\alpha\right),$$

from where the first inequality follows.

Now, from the recursion Eq. (10), we have for a tree ensemble T with root degree $\eta_0 = 1$, and random clause φ assigned to the child of the root,

$$h_T\left(\mathbf{x}_{l+1}^+\right) = \frac{2\,\mathrm{T}_{h_l}\,\varphi^{(1)}\left(\mathbf{s}\right)}{1+\mathrm{T}_{h_l}\,\psi\left(\mathbf{s}\right)}, \qquad \psi\left(s\right) \overset{def}{=} \varphi\left(1,s\right)\varphi\left(-1,s\right)$$

or alternatively,

$$h_T\left(\mathbf{x}_{l+1}^+\right) = \mathrm{T}_{h_l}\,\varphi^{(1)}\left(\mathbf{s}\right) + \left(\mathrm{T}_{h_l}\,\varphi^{(1)}\left(\mathbf{s}\right)\right)\mathcal{G}_k\left(h_l,\mathbf{s}\right), \qquad \mathcal{G}_k\left(h_l,s\right) \overset{def}{=} \left[\frac{1-\mathrm{T}_{h_l}\,\psi\left(s\right)}{1+\mathrm{T}_{h_l}\,\psi\left(s\right)}\right],$$

where $\mathbf{s} \sim \text{Unif}\, S^+\left(\varphi\right)$. Notice that for any antisymmetric function $f\left(s\right)$, we have that $\mathbb{E}_{\mathbf{s}} f\left(\mathbf{s}\right) = \frac{\left(\varphi^{(1)}, f\right)}{\|\varphi\|^2}$. Therefore, due to the fact that $\mathrm{T}_{h_l}\,\varphi^{(1)}\left(s\right)$ is antisymmetric and $\mathcal{G}_k\left(h_l,s\right)$ is symmetric (both in s and h_l, actually), we have the formulas

$$\left\langle h_T\left(\mathbf{x}_{l+1}^+\right)\right\rangle_T = \frac{2}{\|\varphi\|^2}\left\langle\left(\varphi^{(1)}, \frac{\mathrm{T}_{h_l}\,\varphi^{(1)}\left(\mathbf{s}\right)}{1+\mathrm{T}_{h_l}\,\psi\left(\mathbf{s}\right)}\right)\right\rangle_T \tag{14}$$

and

$$\left\langle h_T\left(\mathbf{x}_{l+1}^+\right)\right\rangle_T = \left\langle\frac{\left(\varphi^{(1)}, \mathrm{T}_{h_l}\,\varphi^{(1)}\right)}{\|\varphi\|^2}\right\rangle_T + \left\langle\frac{\left(\varphi^{(1)}, \left(\mathrm{T}_{h_l}\,\varphi^{(1)}\right)\mathcal{G}_k\left(h_l,\cdot\right)\right)}{\|\varphi\|^2}\right\rangle_T. \tag{15}$$

In the last expression, the first term is equal to $\frac{\left(\varphi^{(1)}, \mathrm{T}_{\langle h_l\rangle_T}\,\varphi^{(1)}\right)}{\|\varphi\|^2}$, while the second term can be writen, using Fourier expansion, as

$$\frac{1}{\|\varphi\|^2}\sum_{\substack{Q\subseteq[k-1]\\ |Q|\text{ odd}}}\left(\varphi^{(1)}, \gamma_Q\mathbb{E}_{h_l}\left[\gamma_Q\left(h_l\right)\mathcal{G}_k\left(h_l,\cdot\right)\right]\right)\left(\varphi^{(1)},\gamma_Q\right).$$

Using the fact that $\mathbb{E}\left|\mathbf{X}\right| \leq \left(\mathbb{E}\mathbf{X}\right)^{1/2}$ for consistent random variables, we can bound the terms with $|Q| \geq 3$ by

$$\frac{\left|\left(\varphi^{(1)},1\right)\right|}{\|\varphi\|^2}\sum_{\substack{Q\subseteq[k-1]\\ |Q|\geq 3\text{ odd}}}\left|\left(\varphi^{(1)},\gamma_Q\right)\right|\left(\prod_{i\in Q}\left\langle h_{T_i}\left(\mathbf{x}_l^+\right)\right\rangle_{T_i}\right)^{1/2}.$$

Also, using the fact that for any even function $f\left(x\right)$ with $0 \leq f\left(x\right) \leq 1$ and a consistent random variable $\mathbf{X}$, we have

$$|\mathbb{E}[\mathbf{X}f(\mathbf{X})]| = |\mathbb{E}[2\mathbf{X}^2 f(\mathbf{X})/(1+\mathbf{X})\mathbb{I}_{\{\mathbf{X}\geq 0\}}]| \leq |\mathbb{E}[2\mathbf{X}^2/(1+\mathbf{X})\mathbb{I}_{\{\mathbf{X}\geq 0\}}]| = |\mathbb{E}[\mathbf{X}]|,$$

we can bound the terms with $|Q| = 1$, by

$$\frac{\left|\left(\varphi^{(1)},1\right)\right|}{\|\varphi\|^2}\sum_{i=1}^{k-1}\left(\varphi^{(1)},\gamma_{\{i\}}\right)\left|\left\langle h_{T_i}\left(\mathbf{x}_l^+\right)\right\rangle_{T_i}\right|.$$

Therefore, for a random tCSP (α, p) with root degree $\eta_0 = 1$, we obtain after averaging

$$\widehat{h}_{l+1}^{\text{ave}} \leq \mathbb{E}_\varphi\frac{\left(\varphi^{(1)}, \mathrm{T}_{h_l^{\text{ave}}}\,\varphi^{(1)}\right)}{\|\varphi\|^2} + \mathbb{E}_\varphi\left[\frac{2\,\mathrm{I}_1\left(\varphi\right)}{\|\varphi\|^2}\sum_{\substack{Q\subseteq[k-1]\\ |Q|\geq 3\text{ odd}}}\left|\left(\varphi^{(1)},\gamma_Q\right)\right|\left(\sqrt{h_l^{\text{ave}}}\right)^{\max\{|Q|,2\}}\right],$$

which is precisely the second inequality in the Lemma.

Now, suppose that h_l is supported on non-negative values and let $A_s = \{h_l : \mathrm{T}_{h_l}\,\varphi^{(1)}(s) > 0\}$. Notice that the complement of A_s is $-A_s$ (due to the antisymmetry of $\mathrm{T}_{h_l}\,\varphi^{(1)}(s)$ respect to h_l). Therefore, using the consistency of the random variables $h_{l,i}$, from the Eq. (14) we get

$$\begin{aligned}
\left\langle h_T\left(\mathbf{x}_{l+1}^{+}\right)\right\rangle_T &= \frac{2}{\|\varphi\|^2}\left\langle \left(\varphi^{(1)}, \frac{\mathrm{T}_{h_l}\,\varphi^{(1)}(\mathbf{s})}{1+\mathrm{T}_{h_l}\,\psi(\mathbf{s})}\right)\mathbb{I}(h_l\in A_{\mathbf{s}}) - \left(\varphi^{(1)}, \frac{\mathrm{T}_{-h_l}\,\varphi^{(1)}(\mathbf{s})}{1+\mathrm{T}_{-h_l}\,\psi(\mathbf{s})}\right)\mathbb{I}(-h_l\in A_{\mathbf{s}})\right\rangle_T \\
&= \frac{2}{\|\varphi\|^2}\left\langle \left(\varphi^{(1)}, \frac{\mathrm{T}_{h_l}\,\varphi^{(1)}(\mathbf{s})}{1+\mathrm{T}_{h_l}\,\psi(\mathbf{s})}\right)\mathbb{I}(h_l\in A_{\mathbf{s}})\left[1-\prod_{i=1}^{k-1}\frac{1-h_{l,i}}{1+h_{l,i}}\right]\right\rangle_T \\
&\le \frac{2}{\|\varphi\|^2}\left\langle \left(\varphi^{(1)}, \mathrm{T}_{h_l}\,\varphi^{(1)}(\mathbf{s})\right)\mathbb{I}(h_l\in A_{\mathbf{s}})\left[1-\prod_{i=1}^{k-1}\frac{1-h_{l,i}}{1+h_{l,i}}\right]\right\rangle_T \\
&= \frac{2\left(\varphi^{(1)}, \mathrm{T}_{\langle h_l\rangle_T}\,\varphi^{(1)}(\mathbf{s})\right)}{\|\varphi\|^2}.
\end{aligned}$$

Therefore, for a random tCSP (α, p) with root degree $\eta_0 = 1$, we obtain after averaging, that

$$\widehat{h}_{l+1}^{\mathrm{ave}} \le 2\mathbb{E}_\varphi \frac{\left(\varphi^{(1)}, \mathrm{T}_{h_l^{\mathrm{ave}}}\,\varphi^{(1)}\right)}{\|\varphi\|^2},$$

which corresponds to the last inequality of the lemma. □

We now return to completing the proof of Theorem 4.1.

Proof of the lower bound in Theorem 4.1. If $\theta = 1$, T_1 is the identity operator whence $(\varphi^{(1)}, \mathrm{T}_1\varphi^{(1)}) = \mathrm{I}_1(\varphi)$. We have therefore $F_k(1) = 1/\Omega_k$. Now, expanding in Fourier series we get,

$$(\varphi^{(1)}, \mathrm{T}_\theta\,\varphi^{(1)}) = \sum_{Q\subseteq[k-1]} |\left(\varphi^{(1)}, \gamma_Q\right)|^2\,\theta^{|Q|} = \sum_{Q\subseteq[k], Q\ni\{i\}} |(\varphi, \gamma_Q)|^2\;\theta^{|Q|-1}\,.$$

By the *Fourier expansion condition*,

$$F_k(\theta) \le e^{-Ck(1-\theta)}/\Omega_k. \tag{16}$$

Now fix $\alpha = (1-\delta)(\Omega_k\log k)/k$, whence, by Lemma 4.3, $h_1^{\mathrm{ave}} \le 1 - k^{-1+\delta}$, and h_1 is supported on non-negative reals. Using Eq. (13), we get $\widehat{h}_2^{\mathrm{av}} \le e^{-Ck^\delta}/\Omega_k$, and therefore,

$$h_2^{\mathrm{av}} \le 1 - \exp\{-2(1-\delta)e^{-Ck^\delta}\log k\,\} \le e^{-Ck^\delta/2}\,.$$

On the other hand, from the Eq. (3), we obtain the following bounds for $F_k(\theta)$, $R_k(\theta)$:

$$F_k(\theta) \le 2\mathbb{E}_\varphi\left[\frac{\sum_{i=1}^{k-1}|\left(\varphi^{(1)}, \gamma_{\{i\}}\right)|^2}{\|\varphi\|^2}\right]\theta + 2\mathbb{E}_\varphi\left[\frac{\mathrm{I}_1(\varphi)}{\|\varphi\|^2}\right]\theta^2 \le \left(Ae^{-Ck/2}\theta + \theta^2\right)/\Omega_k.$$

On the other hand,

$$R_k(\theta) \le 2\mathbb{E}_{\varphi\mathbf{i}}\left[\frac{2\,\mathrm{I}_1(\varphi)}{\|\varphi\|^2}\sum_{i=1}^{k-1}|\left(\varphi^{(1)}, \gamma_{\{i\}}\right)|^2\right]\theta^2 + 2\mathbb{E}_\varphi\left[\frac{2\,\mathrm{I}_1(\varphi)}{\|\varphi\|^2}\sum_{Q\subseteq[k-1]}|\left(\varphi^{(1)}, \gamma_Q\right)|\right]\theta^3 \le (Ae^{-Ck/2}\theta^2 + k^a\theta^3)/\Omega_k\,,$$

Therefore, for all ℓ we have

$$h_{\ell+1}^{\mathrm{av}} \le 1 - e^{-k\alpha[F_k(h_\ell^{\mathrm{av}})+R_k(h_\ell^{\mathrm{av}})]} \le (1-\delta)\log k(2Ae^{-Ck/2}h_\ell^{\mathrm{av}} + 2k^a(h_\ell^{\mathrm{av}})^{3/2})\,.$$

which implies $h_\ell^{\mathrm{av}} \to 0$ if, for some $\ell > 0$, $h_\ell^{\mathrm{av}} \le k^{-5a}$, thus finishing the proof. □

5 Reconstruction on Trees to Graphs: the case of proper q colorings

In this section we prove that the set of solutions of the proper q-coloring ensemble satisfies the *sphericity* condition described in the section 3.3.

Given two assignments $\underline{x}^{(1)}$, $\underline{x}^{(2)}$ of the variables $x_1, \ldots, x_n$, their joint type $v_{\underline{x}^{(1)},\underline{x}^{(2)}}$ is the $q \times q$ matrix with $v_{\underline{x}^{(1)},\underline{x}^{(2)}}(i,j) \stackrel{def}{=} \frac{1}{n}\#\left\{t \in G : \underline{x}^{(1)}(t) = i \text{ and } \underline{x}^{(2)}(t) = j\right\}$. We consider random assignments $\underline{\mathbf{x}}^{(1)}$, $\underline{\mathbf{x}}^{(2)}$ taken uniformly and independently over all the satisfying assignments of a random instance of the q-coloring model with edge-variable density α. Our purpose is to prove that for all $\delta > 0$, $||v_{\underline{\mathbf{x}}^{(1)},\underline{\mathbf{x}}^{(2)}} - \overline{\nu}||_{\rm TV} \leq \delta$ w.h.p., where $\overline{v}$ is the matrix with all entries equal to $1/q^2$.

Our argument makes crucial use of the following estimate for the partition function from [AC08].

Lemma 5.1 ([AC08, Lemma 7]) *Let Z be the number of satisfying assignments of a random instance of the q-coloring model with edge-variable density $\alpha < q\log q$, then*

$$\mathbf{E}Z \geq \Omega\left(\frac{1}{n^{(q-1)/2}}\right)\left[q\left(1-\frac{1}{q}\right)^{\alpha}\right]^{n},$$

and, for some function $f(n)$ of order $o(n)$, we have $\text{Prob}\left(Z < e^{-f(n)}\mathbf{E}[Z]\right) \to 0$ *as $n \to \infty$.*

Let us introduce some notation. If w is a vector of lenght q and v is a $q \times q$ matrix v, let $\mathcal{H}$ and $\mathcal{E}$ denote their entropy an their enrgy respectively, where

$$\mathcal{H}(v) = -\sum_{i,j} v(i,j)\log v(i,j)\ , \quad \mathcal{H}(w) = -\sum_{i} w(i)\log w(i)$$

$$\mathcal{E}(v) = \log\left(1 - \sum_i\left(\sum_j v(i,j)\right)^2 - \sum_j\left(\sum_i v(i,j)\right)^2 + \sum_{i,j} v(i,j)^2\right)\ , \quad \mathcal{E}(w) = \log\left(1 - \sum_i w(i)^2\right)$$

Let $\mathcal{B}_q^{\epsilon}$ consists of all the q-vectors w with nonegative entries such that $\sum_i w(i) = 1$ and $\|w - \overline{w}\|^2 > \epsilon$. Similarly, let $\mathcal{B}_{q\times q}^{\delta,\epsilon}$ be the set of all the $q \times q$ matrices with nonegative entries such that $\|(v-\overline{v})1\|^2 \leq \delta$, $\|1^t(v-\overline{v})\|^2 \leq \delta$ and $\|v - \overline{v}\|^2 \geq \epsilon$.

Our goal in this section is to prove the following theorem.

Theorem 5.2 *Let $\underline{\mathbf{x}}^{(1)}$, $\underline{\mathbf{x}}^{(2)}$ be random assignments taken uniformly and independently over all the satisfying assignments of a random instance of the q-coloring model with edge-variable density α. If $\alpha < (q-1)\log(q-1)$, then for any $\epsilon > 0$,*

$$\text{Prob}\left(\left\|v_{\underline{\mathbf{x}}^{(1)},\underline{\mathbf{x}}^{(2)}} - \overline{v}\right\|^2 > \epsilon\right) \to 0 \text{ as } n \to \infty.$$

We will present several lemmas before returning to the proof of the Theorem. First we introduce estimations concerning an additive functional depending on the energy and entropy of a vector of lenght q.

Lemma 5.3 *If $w \in \mathcal{B}_q^{\epsilon}$, then $\mathcal{H}(w) + \alpha\mathcal{E}(w) \leq [\mathcal{H}(\overline{w}) + \alpha\mathcal{E}(\overline{w})] - \frac{\alpha\epsilon}{2(1-1/q)}$.*

Proof. Notice that $[\mathcal{H}(\overline{w}) + \alpha\mathcal{E}(\overline{w})] - [\mathcal{H}(w) + \alpha\mathcal{E}(w)] = \alpha\log\left(\frac{1-\|\overline{w}\|^2}{1-\|w\|^2}\right)$. This quantity is bounded below by $\alpha\log\left(1 + \frac{\epsilon}{1-1/q}\right)$, and therefore by $\frac{\alpha\epsilon}{2(1-1/q)}$. □

Lemma 5.4 *Let $\underline{\mathbf{x}}$ be a random assignment of the variables taken uniformly over all the satisfying assignments of a random instance of the q-coloring model with edge-variable density $\alpha < q\log q$. Then, for any $\epsilon > 0$,*

$$\text{Prob}\left(\left\|w_{\underline{\mathbf{x}}} - \overline{w}\right\|^2 > \epsilon\right) \to 0 \text{ as } n \to \infty$$

where w is the vector with q entries such that $w_{\underline{\mathbf{x}}}(i) = \frac{1}{n}\#\{v \in G : \underline{\mathbf{x}}_v = i\}$ and $\overline{w}$ is the vector with all entries equal to $1/q$.

Proof. Given a property P, denote by $Z(P)$, the number of satisfying assignments for shich P holds. Choose ξ such that $\xi < \frac{\alpha\epsilon}{2(1-1/q)}$. We have that

$$\text{Prob}\left(\|w_{\underline{\mathbf{x}}} - \overline{w}\|^2 > \epsilon\right) = \mathbf{E}\left[Z\left(\|w_{\underline{\mathbf{x}}}\|^2 > \epsilon + 1/q\right)/Z\right],$$

an expression that we can bound by

$$\frac{\mathbf{E}\left[Z\left(\|w_{\underline{\mathbf{x}}}\|^2 > \epsilon + 1/q\right)\right]}{e^{-n\xi}\mathbf{E}[Z]} + \text{Prob}\left(Z < e^{-n\xi}\mathbf{E}[Z]\right).$$

Now, according to the Lemma 5.1, $\text{Prob}\left(Z < e^{-n\xi}\mathbf{E}[Z]\right) \to 0$, and therefore it is enough to show that the term $\mathbf{E}\left[Z\left(\|w_{\underline{\mathbf{x}}}\|^2 > \epsilon + 1/q\right)\right]/e^{-n\xi}\mathbf{E}[Z]$ vanishes.

Denote by $\mathcal{G}_\epsilon$ the set of all vectors ℓ, with nonegative integer entries, such that $\sum_{i=1}^{q}(\ell_i/n) = 1$ and $\sum_{i=1}^{q}(\ell_i/n)^2 > \epsilon + 1/q$, and denote by Ω_w the set of assignments $\underline{x}$ such that $w_{\underline{x}}$ is equal to the vector w. Now,

$$\begin{aligned}
\mathbf{E}\left[Z\left(\|w_{\underline{\mathbf{x}}}\|^2 > \epsilon + 1/q\right)\right] &= \sum_{\ell\in\mathcal{G}_\epsilon}\sum\nolimits_{\underline{x}\in\Omega_{\ell/n}} \text{Prob}\left(\underline{x} \text{ is a satisfying assignment}\right) \qquad (17)\\
&= \sum_{\ell\in\mathcal{G}_\epsilon}\frac{n!}{\prod_{i=1}^{q}\ell_i!}\left(\left[\frac{n}{n-1}\right]\left[1-\sum_{i=1}^{q}(\ell_i/n)^2\right]\right)^{\alpha n}\\
&\le \sum_{\ell\in\mathcal{G}_\epsilon} 3q^{2q}\sqrt{n}\exp\left(n\left[\mathcal{H}(\ell/n) + c_n\mathcal{E}(\ell/n)\right]\right)\\
&\le 3q^{2q}\sqrt{n}\,|\mathcal{G}_\epsilon|\sup_{\ell\in\mathcal{G}_\epsilon}\left\{\exp\left(n\left[\mathcal{H}(\ell/n) + c_n\mathcal{E}(\ell/n)\right]\right)\right\}.
\end{aligned}$$

Here $|\mathcal{G}_\epsilon|$ is the number of elements of $\mathcal{G}_\epsilon$, which is bounded by n^q. Notice also that if $\ell \in \mathcal{G}_\epsilon$, then $\ell/n \in \mathcal{B}_q^\epsilon$, so that by Lemma 5.3,

$$\begin{aligned}
\mathcal{H}(\ell/n) + \alpha\mathcal{E}(\ell/n) &\le [\mathcal{H}(j_q) + \alpha\mathcal{E}(j_q)] - \frac{\alpha\epsilon}{2(1-1/q)} \qquad (18)\\
&= \log q + \alpha\log(1-1/q) - \frac{\alpha\epsilon}{2(1-1/q)}.
\end{aligned}$$

On the other hand by the Lemma 5.1, there is some constant C such that

$$e^{-n\xi}\mathbf{E}[Z] \ge \frac{C}{n^{(q-1)/2}}e^{-n\xi}\left[q\left(1-\frac{1}{q}\right)^{\alpha}\right]^n. \qquad (19)$$

Combining Eq. (17), (18) and (19), we have that for a polynomial $p(n)$ of degree $3q/2$,

$$\frac{\mathbf{E}\left[Z\left(\|w_{\underline{\mathbf{x}}} - \overline{w}\|^2 > \epsilon\right)\right]}{e^{-n\xi}\mathbf{E}[Z]} \le p(n)\exp\left(n\left[\xi - \frac{\alpha\epsilon}{2(1-1/q)}\right]\right). \qquad (20)$$

From (20), it is now clear that $\frac{\mathbf{E}\left[Z\left(\|w_{\underline{\mathbf{x}}}-\overline{w}\|^2>\epsilon\right)\right]}{e^{-n\xi}\mathbf{E}[Z]} \to 0$ as $n \to \infty$, due to the fact that $\xi - \frac{\alpha\epsilon}{2(1-1/q)} < 0$. □

Next, our objective is to work with the quantity $\kappa_q^{\delta,\epsilon}$, which we define as the upper limit of the interval (indeed, easy to see that this is an interval) consisting of the values c such that

$$\sup_{v\in\mathcal{B}_{q\times q}^{\delta,\epsilon}} \mathcal{H}(v) + c\mathcal{E}(v) \le \mathcal{H}(\overline{v}) + \alpha\mathcal{E}(\overline{v}).$$

To motivate, let us recall that an important part of the second moment argument of Achlioptas and Naor [AN05, Theorem 7] (in showing that the chromatic number $\chi[G(n,d/n)]$ concentrated on two possible values), relied on an optimization of the expression $\mathcal{H}(v)+\alpha\mathcal{E}(v)$ over the Birkoff polytope $\mathcal{B}_{q\times q}$ of the $q\times q$ doubly stochastic matrices. In particular, they proved that, as long as $\alpha\le(q-1)\log(q-1)$, one has

$$\sup_{v\in\mathcal{B}_{q\times q}} \mathcal{H}(v)+\alpha\mathcal{E}(v)=\mathcal{H}(\overline{v})+\alpha\mathcal{E}(\overline{v})\,. \tag{21}$$

Since $\mathcal{B}^{0,\epsilon}_{q\times q}\subseteq\mathcal{B}_{q\times q}$, we have $\kappa_q^{0,\epsilon}\ge(q-1)\log(q-1)$. The next lemma says that $\sup_{v\in\mathcal{B}^{\delta,\epsilon}_{q\times q}} \mathcal{H}(v)+\alpha\mathcal{E}(v)$ is in fact 'separated' from $\mathcal{H}(\overline{v})+\alpha\mathcal{E}(\overline{v})$, provided that $\alpha<\kappa_q^{\delta,\epsilon}$.

Lemma 5.5 *Suppose that $v\in\mathcal{B}^{\delta,\epsilon}_{q\times q}$ where $\epsilon>2\delta$, then, if $\alpha<\kappa_q^{\delta,\epsilon}$, we have that*

$$[\mathcal{H}(v)+\alpha\mathcal{E}(v)]\le[\mathcal{H}(\overline{v})+\alpha\mathcal{E}(\overline{v})]-\frac{\left(\kappa_q^{\delta,\epsilon}-\alpha\right)}{2\left(1-1/q\right)^2}\left[\epsilon-2\delta\right]\,.$$

Proof. Indeed,

$$\begin{aligned}
[\mathcal{H}(\overline{v})+\alpha\mathcal{E}(\overline{v})]-[\mathcal{H}(v)+\alpha\mathcal{E}(v)] &= \left[\mathcal{H}(\overline{v})+\kappa_q^{\delta,\epsilon}\mathcal{E}(\overline{v})\right]-\left[\mathcal{H}(v)+\kappa_q^{\delta,\epsilon}\mathcal{E}(v)\right]+\left(\kappa_q^{\delta,\epsilon}-\alpha\right)\left[\mathcal{E}(v)-\mathcal{E}(\overline{v})\right]\\
&\ge\left(\kappa_q^{\delta,\epsilon}-\alpha\right)\left[\log\left(1+\frac{1}{\left(1-1/q\right)^2}\left[\|v-\overline{v}\|^2-\|(v-\overline{v})\,1\|^2-\left\|1^t\,(v-\overline{v})\right\|^2\right]\right)\right]\\
&\ge\frac{\left(\kappa_q^{\delta,\epsilon}-\alpha\right)}{2\left(1-1/q\right)^2}\left[\epsilon-2\delta\right].
\end{aligned}$$

□

Lemma 5.6 *Given $\epsilon>0$ and $\alpha<\alpha_q=(q-1)\log(q-1)$, there exists $\delta>0$ such that $\kappa_q^{\delta,\epsilon}\ge\alpha$.*

Proof. Assume the contrary, then there exists a sequence $\delta_n\downarrow 0$ such that $\kappa_q^{\delta_n,\epsilon}<\alpha$ for each n. Due to the continuity of $\exp(\mathcal{H}(v)+\alpha\mathcal{E}(v))$ in the compact set $\mathcal{B}^{\delta,\epsilon}_{q\times q}$, the supremum of $\exp(\mathcal{H}(v)+\alpha_q\mathcal{E}(v))$ is reached at a matrix $v_{\delta_n}\in\mathcal{B}^{\delta_n,\epsilon}_{q\times q}\subseteq\mathcal{P}_{q\times q}$, and due to the compactness of $\mathcal{P}_{q\times q}$, a subsequence $\left\{v_{\delta_{n_k}}\right\}_{k\ge1}$ of these matrices converges in $\mathcal{P}_{q\times q}$ to a matrix $v\in\mathcal{B}^{0,\epsilon}_{q\times q}$. Therefore $\mathcal{H}(v)+\alpha\mathcal{E}(v)\le\mathcal{H}(\overline{v})+\alpha\mathcal{E}(\overline{v})-\frac{(\alpha_q-\alpha)\epsilon}{2(1-1/q)^2}$. On the other hand,

$$\mathcal{H}(v)+\alpha\mathcal{E}(v))\ge\liminf_{k\to\infty}\mathcal{H}(v_{\delta_{n_k}})+\alpha\mathcal{E}\left(v_{\delta_{n_k}}\right)\ge\mathcal{H}(\overline{v})+\alpha\mathcal{E}\left(\overline{v}\right),$$

obtaining a contradiction. □

Proof of Theorem 5.2. Given a property P, denote by $Z^{(2)}(P)$, the number of pairs of satisfying assignments for which P holds. Take α' such that $\alpha<\alpha'<(q-1)\log(q-1)$ and use Lemma 5.6 to choose δ such that $\kappa_q^{\delta,\epsilon}\ge\alpha'$, guaranteeing also that $2\delta<\epsilon$. Now, let ξ be a positive real such that $2\xi<\frac{(\alpha'-\alpha)}{2(1-1/q)^2}\left[\epsilon-2\delta\right]$. We have that

$$\operatorname{Prob}\left(\left\|v_{\underline{\mathbf{x}}^{(1)},\underline{\mathbf{x}}^{(2)}}-\overline{v}\right\|^2>\epsilon\right)=\mathbf{E}\left[Z^{(2)}\left(\left\|v_{\underline{\mathbf{x}}^{(1)},\underline{\mathbf{x}}^{(2)}}-\overline{v}\right\|^2>\epsilon\right)/Z^2\right],$$

which is bounded by the addition of the terms $E\left[Z^{(2)}\left(v_{\underline{\mathbf{x}}^{(1)},\underline{\mathbf{x}}^{(2)}}\in\mathcal{B}^{\delta,\epsilon}_{q\times q}\right)\right]/e^{-2n\xi}\mathbf{E}\left[Z\right]^2$, $\operatorname{Prob}\left(Z<e^{-n\xi}\mathbf{E}\left[Z\right]\right)$, $\operatorname{Prob}\left(\left\|\left(v_{\underline{\mathbf{x}}^{(1)},\underline{\mathbf{x}}^{(2)}}-\overline{v}\right)1\right\|^2>\epsilon\right)$ and $\operatorname{Prob}\left(\left\|1^t\left(v_{\underline{\mathbf{x}}^{(1)},\underline{\mathbf{x}}^{(2)}}-\overline{v}\right)\right\|^2>\epsilon\right)$. Now, Lemma 5.1 implies that the second term vanishes and lemma 5.4 implies that the last two terms go to zero. Therefore, to show that $\operatorname{Prob}\left(\left\|v_{\underline{\mathbf{x}}^{(1)},\underline{\mathbf{x}}^{(2)}}-\overline{v}\right\|^2>\epsilon\right)\to0$ is sufficient to prove that the term $\mathbf{E}\left[Z^{(2)}\left(v_{\underline{\mathbf{x}}^{(1)},\underline{\mathbf{x}}^{(2)}}\in\mathcal{B}^{\delta,\epsilon}_{q\times q}\right)\right]/e^{-2n\xi}\mathbf{E}\left[Z\right]^2$ vanishes.

Denoting by $\mathcal{G}_{\epsilon,\delta}$ the set of all $q\times q$ matrices L, with nonnegative integer entries, such that $L/n \in \mathcal{B}^{\delta,\epsilon}_{q\times q}$, and denoting by Ω_v the set of pairs of colorings x_1, x_2 such that v_{x_1,x_2} is equal to the matrix v, we have

$$\begin{aligned}\mathbf{E}\left[Z^{(2)}\left(v_{\underline{\mathbf{x}}^{(1)},\underline{\mathbf{x}}^{(2)}} \in \mathcal{B}^{\delta,\epsilon}_{q\times q}\right)\right] &= \sum_{L\in\mathcal{G}_{\epsilon,\delta}} \sum\nolimits_{x_1,x_2\in\Omega_{L/n}} \operatorname{Prob}\left(x_1 \text{ and } x_2 \text{ are satisfying assignments}\right) \\ &= \sum_{L\in\mathcal{G}_\epsilon} \frac{n!}{\prod\limits_{i,j} L_{ij}!}\left[\frac{n}{n-1}\right]^{\alpha n}\left(1-\sum_i\left(\sum_j L_{ij}/n\right)^2-\sum_j\left(\sum_i L_{ij}/n\right)^2+\sum_{i,j}\left(L_{ij}/n\right)^2\right)^{\alpha n} \\ &\leq \sum_{L\in\mathcal{G}_{\epsilon,\delta}} 3q^{2q}\sqrt{n}\exp\left(n\left[\mathcal{H}\left(L/n\right)+\alpha E\left(L/n\right)\right]\right).\end{aligned}$$

And now, because $\kappa^{\delta,\epsilon}_q \geq \alpha' > \alpha$ and $L/n \in \mathcal{B}^{\delta,\epsilon}_{q\times q}$ where $2\delta < \epsilon$, we can invoke Lemma 5.5 to get that

$$[\mathcal{H}(L/n)+\alpha\mathcal{E}(L/n)] \leq [\mathcal{H}(\overline{v})+\alpha\mathcal{E}(\overline{v})] - \frac{(\alpha'-\alpha)}{2\left(1-1/q\right)^2}\left[\epsilon-2\delta\right] .$$

Therefore,

$$\mathbf{E}\left[Z^{(2)}\left(v_{\underline{\mathbf{x}}^{(1)},\underline{\mathbf{x}}^{(2)}} \in \mathcal{B}^{\delta,\epsilon}_{q\times q}\right)\right] \leq 3q^{2q}\sqrt{n}\left|\mathcal{G}_{\epsilon,\delta}\right|\left[q\left(1-1/q\right)^\alpha\right]^{2n}\exp\left(-n\frac{(\alpha'-\alpha)}{2\left(1-1/q\right)^2}\left[\epsilon-2\delta\right]\right),$$

where $|\mathcal{G}_{\epsilon,\delta}|$ is the number of elements in $\mathcal{G}_{\epsilon,\delta}$, which is bounded by n^{q^2}. On the other hand by Lemma 5.1, we have that for some constant C,

$$e^{-2n\xi}\mathbf{E}\left[Z\right]^2 \geq \frac{C}{n^{(q-1)}}e^{-2n\xi}\left[q\left(1-\frac{1}{q}\right)^\alpha\right]^{2n} .$$

Hence, for a polynomial $p\left(n\right)$ of degree q^2+q-1, we have

$$\frac{\mathbf{E}\left[Z^{(2)}\left(v_{\underline{\mathbf{x}}^{(1)},\underline{\mathbf{x}}^{(2)}} \in \mathcal{B}^{\delta,\epsilon}_{q\times q}\right)\right]}{e^{-2n\xi}\mathbf{E}\left[Z\right]^2} \leq p(n)\exp\left\{n\left(2\xi-\frac{(\alpha'-\alpha)}{2\left(1-1/q\right)^2}\left[\epsilon-2\delta\right]\right)\right\}.$$

Due to the fact that $2\xi < \frac{(\alpha'-\alpha)}{2(1-1/q)^2}\left[\epsilon-2\delta\right]$, it is now clear that $\frac{\mathbf{E}\left[Z^{(2)}\left(v_{\underline{\mathbf{x}}^{(1)},\underline{\mathbf{x}}^{(2)}}\in\mathcal{B}^{\delta,\epsilon}_{q\times q}\right)\right]}{e^{-2n\xi}\mathbf{E}[Z]^2} \to 0$ as $n\to\infty$. ☐

Acknowledgments. The last two authors are grateful to Eric Vigoda and Linji Yang for many insightful discussions on reconstruction problems, and for their role in the early development of this project. The authors also gratefully acknowledge the support and the hospitality of BIRS (Canada) and DIMACS (USA), which provided ideal environs for carrying out a significant part of this research collaboration.

References

[AC08] D. Achlioptas and A. Coja-Oghlan, *Algorithmic Barriers from Phase Transitions,* Proc. of IEEE FOCS 2008.

[AM02] D. Achlioptas and C. Moore, *The asymptotic order of the random k-SAT threshold,* Proc. of IEEE FOCS 2002.

[AN05] D. Achlioptas and A. Naor, *The two possible values of the chromatic number of a random graph,* Annals of Mathematics, **162** (2005), 1333–1349.

[ANP05] D. Achlioptas, A. Naor, and Y. Peres, *Rigorous location of phase transitions in hard optimization problems*, Nature **435** (2005), 759–764.

[AR06] D. Achlioptas and F. Ricci-Tersenghi, *On the solution-space geometry of random constraint satisfaction problems,* Proc. of ACM STOC 2006.

[AS04] D. Aldous, J. M. Steele, 'The Objective Method: Probabilistic Combinatorial Optimization and Local Weak Convergence,' in *Probability on discrete structures*, H. Kesten (ed.), New York, 2004.

[BK+05] N .Berger, C. Kenyon, E. Mossel and Y. Peres, *Glauber dynamics on trees and hyperbolic graphs,* Probab. Theory Relat. Fields, **131** (2005) 311-340.

[BVV07] N. Bhatnagar, J. Vera, and E. Vigoda. *Reconstruction for colorings on trees.* http://front.math.ucdavis.edu/0711.3664, 2007.

[Geo88] H.-O. Georgii. 'Gibbs Measures and Phase Transitions,' de Gruyter, Berlin, 1988.

[GM07] A. Gerschenfeld, A. Montanari. *Reconstruction for models on random graphs*, Proc. of IEEE FOCS 2007.

[HPT08] J. Hartigan, D. Pollard, S. Tatikonda. *Conditioned Poisson Distributions and the concentration of chromatic numbers.* http://www.stat.yale.edu/~pollard/Papers/chromatic.30june08.pdf

[KM+07] F. Krzakala, A. Montanari, F. Ricci-Tersenghi, G. Semerjian and L. Zdeborova, *Gibbs States and the Set of Solutions of Random Constraint Satisfaction Problems,* Proc. Natl. Acad. Sci. 1004, 10318 (2007)

[MPZ02] M. Mézard, G. Parisi and R. Zecchina *Analytic and Algorithmic Solution of Random Satisfiability Problems,* Science **297** (2002), 812-815.

[MRZ03] M. Mézard, F. Ricci-Tersenghi and R. Zecchina *Alternative solutions to diluted p-spin models and XORSAT problems,* J. Stat. Phys. **111** (2003), 505-53

[MZ02] M. Mézard and R. Zecchina *Random K-satisfiability problem: From an analytic solution to an efficient algorithm,* Phys. Rev. E **66** (2002), 056126

[MM06] M. Mézard and A. Montanari, *Reconstruction on Trees and Spin Glass Transition,* J. Stat. Phys. **124** (2006), 1317-1350.

[MM09] M. Mézard and A. Montanari, 'Information, Physics, and Computation,' Oxford University Press, Oxford, 2009.

[MMZ05] M. Mézard, T. Mora and R. Zecchina, *Clustering of Solutions in the Random Satisfiability Problem,* Phys. Rev. Lett. **94** (2005), 197205.

[CD+03] S. Cocco, O. Dubois, J. Mandler, R. Monasson, *Rigorous Decimation-Based Construction of Ground States for Spin-Glass Models on Random Lattices*, Phys. Rev. Lett. **90** (2003), 047205

[MP03] E. Mossel and Y. Peres, *Information Flow on Trees,* Ann. Appl. Probab. **13** (2003), 817-844.

[RU08] T. Richardson and R. Urbanke, 'Modern Coding Theory', Cambridge University Press, Cambridge, UK, 2008

[Sem08] G. Semerjian. *On the freezing of variables in random constraint satisfaction problems.* J. Stat. Phys., 130-251, 2008.

[Sly08] A. Sly. *Reconstruction of random colourings.* http://front.math.ucdavis.edu/0802.3487, 2008.

A Constrained partition function for binary CSP's

In this section, we prove Proposition 3.1. Given a random CSP(n, p, α) ensemble $\{\varphi_a\}_{a=1}^{\alpha n}$, consider the statistic $L_n(\varphi) = \frac{1}{\alpha n}\#\{a : \varphi_a = \varphi\}$, and denote by CSP$(n, p, \alpha; \widetilde{p}_n)$ the ensemble $\{\varphi_a\}_{a=1}^{\alpha n}$ conditioned on $L_n = \widetilde{p}_n$.. Also, denote by $\overline{\text{CSP}}(n, p, \alpha)$ the ensemble $\{\varphi_a\}_{a=1}^{\alpha n}$ conditioned on $\|L_n - p\|_{TV} < 1/n^{1/2-\gamma}$, where γ is a fixed positive constant. Because Prob$\left(\|L_n - p\|_{TV} \geq 1/n^{1/2-\gamma}\right)$ goes to zero (by the central limit theorem), the probability measures induced by CSP(n, p, α) and $\overline{\text{CSP}}(n, p, \alpha)$ become equivalent as $n \to \infty$.

A binary configuration $\underline{x}$ is said to be balanced if $|\underline{x} \cdot \underline{1}| \leq 1$. We will use Z and Z_b, to denote the variable that counts the number of satisfying assignments and balanced satisfying assignments, respectively, of a random CSP ensemble. Given two binary assignments $\underline{x}^{(1)}, \underline{x}^{(2)}$, we define their overlap as $Q_{12} \stackrel{def}{=} \underline{x}^{(1)} \cdot \underline{x}^{(2)}/n = \sum_{i=1}^n x_i^{(1)} x_i^{(2)}/n$. In other words $(1 - Q_{12})/2$ is the normalized Hamming distance of $\underline{x}^{(1)}$ and $\underline{x}^{(2)}$.

The upper bound in Proposition 3.1 follows from a first moment calculation. In fact, for a random $\overline{\text{CSP}}(n, p, \alpha)$, we have

$$\text{Prob}\,(Z = 0) \leq \mathbf{E}\,[Z] = \sum_{x \in \{-1,1\}^k} \text{Prob}\,(x \text{ is a satisfying assignment}) = \sum_{\frac{\underline{x}\cdot\underline{1}}{n} = \theta} \prod_{\varphi} \|\varphi\|_\theta^{2L_n(\varphi)\alpha n}$$
$$\leq \exp\left(n\left\{\log 2 + \alpha \sum_{\varphi} p(\varphi) \log \|\varphi\|^2 + \mathrm{O}\left(1/n^{1/2-\gamma}\right)\right\}\right),$$

and the last quantity goes to zero whenever $\alpha > (1+\epsilon)\,\widehat{\Omega}_k \log 2$.

To establish the corresponding lower bound, we use the second moment method, but first we need two lemmas.

Lemma A.1 *Given a random* CSP$(n, p, \alpha; \widetilde{p}_n)$ *ensemble, let* $Z_{\text{b}}(|Q_{12}| \geq \delta)$ *be the number of balanced solution pairs* $\underline{x}^{(1)}$, $\underline{x}^{(2)} \in \{+1, -1\}^n$ *with overlap larger than* δ. *Then,*

$$\frac{\mathbf{E}\,[Z_{\text{b}}(|Q_{12}| \geq \delta)]}{[\mathbf{E}Z_{\text{b}}]^2} \leq n \,\exp\left\{n\left[\sup_{\theta \geq \delta} \Phi(\theta)\right]\right\},$$

where

$$\Phi(\theta) \stackrel{def}{=} H(\theta) + \alpha \mathbb{E}_{\varphi \sim \widetilde{p}_n} \log \left\{\frac{(\varphi, \mathrm{T}_\theta\, \varphi)}{\|\varphi\|^4}\right\},$$

and $H(\theta) \equiv -\frac{1+\theta}{2}\log(1+\theta) - \frac{1-\theta}{2}\log(1-\theta)$.

Proof. For simplicity take n to be even. Let φ be a boolean function, and let $\pi : [k] \to [n]$ be a uniform random assignation for the variables in φ. Now, given two *balanced* vectors $\underline{x}^{(1)}$, $\underline{x}^{(2)} \in \{-1, 1\}^n$, we have

$$\mathbb{E}_\pi\left[\varphi\left(\underline{x}_{\pi_1}^{(1)}, \ldots \underline{x}_{\pi_k}^{(1)}\right)\varphi\left(\underline{x}_{\pi_1}^{(2)}, \ldots, \underline{x}_{\pi_k}^{(2)}\right)\right] = (\varphi, \mathrm{T}_\theta\, \varphi),$$

where $\theta = Q_{12}$. Therefore, for some constant $C > 0$,

$$\mathbb{E}_{\pi_a} Z_{\text{b}}(|Q_{12}| \geq \delta) = \sum_{\theta \geq \delta} \sum_{Q_{12} = \theta} \prod_{\varphi} |(\varphi, \mathrm{T}_\theta\, \varphi)|^{L_n(\varphi)\alpha n}$$
$$< \sum_{\theta \geq \delta} \frac{C}{n^{3/2}} \exp\left(n\left\{\mathcal{H}\left(\frac{1+\theta}{4}, \frac{1+\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}\right) + \alpha \sum_{\varphi} L_n(\varphi) \log(\varphi, \mathrm{T}_\theta\, \varphi)\right\}\right).$$

where $\mathcal{H}(\cdot)$ is the entropy function. On the other hand, for some positive C',

$$\mathbb{E}_{\pi_a} Z_{\text{b}} = \sum_{\underline{x} \text{ balanced}} \prod_{\varphi} \|\varphi\|^{2L_n(\varphi)\alpha n}$$
$$> \frac{C'}{n^{1/2}} \exp\left(n\left\{\mathcal{H}\left(\frac{1}{2}, \frac{1}{2}\right) + \alpha \sum_{\varphi} L_n(\varphi) \log \|\varphi\|^2\right\}\right).$$

It is straightforward now to check that

$$\frac{\mathbf{E} Z_{\rm b}(|Q_{12}| \geq \delta)}{\mathbf{E}\left(Z_{\rm b}\right)^2} < \sum_{\theta \geq \delta} \frac{C''}{n^{1/2}} \exp\left(n\left[\Phi\left(\theta\right)\right]\right) \tag{22}$$

and therefore $\frac{\mathbf{E}\, Z_{\rm b}(|Q_{12}| \geq \delta)}{(\mathbf{E} Z_{\rm b})^2} < n \exp\left(n\left[\sup_{\theta \geq \delta} \Phi\left(\theta\right)\right]\right)$. □

Lemma A.2 *Given a random* CSP$(n, p, \alpha; \widetilde{p})$ *ensemble, if* $\alpha \leq (1-\varepsilon)\Omega_{k,\widetilde{p}_n} \log 2$*, where* $\frac{1}{\Omega_{k,\widetilde{p}_n}} \stackrel{def}{=} \mathbb{E}_{\varphi \sim \widetilde{p}_n} \frac{2\,\mathrm{I}_1(\varphi)}{\|\varphi\|^2}$*, then for any* $\delta > 0$ *there exists* $C(\delta, \varepsilon) > 0$ *such that*

$$\mathbf{E}\left[Z_{\rm b}(|Q_{12}| \geq \delta)\right] \leq e^{-n[C(\delta,\epsilon)]} \left(\mathbf{E} Z_{\rm b}\right)^2 .$$

Moreover, as $\delta \to 0$*,* $C(\delta, \epsilon) = \Omega\left(\delta^2\right)$*.*

Proof. In view of the previous lemma, it is sufficient to prove that the function $\theta \mapsto \Phi(\theta)$ achieves its maximum over the interval $[0, 1]$ uniquely at $\theta = 0$. To establish the second statement, then it will be enough to prove that $-\Phi\left(\theta\right) = \Omega\left(\theta^2\right)$ as $\theta \to 0$.

Fix $\alpha \leq (1-\varepsilon)\Omega_k \log 2 \leq (1-\varepsilon)\widehat{\Omega}_k \log 2$. We will prove the thesis claim by considering three different regimes for θ: $0 < \theta \leq e^{-ak}$, $e^{-ak} \leq \theta \leq 1 - \varepsilon^{1/2}$ and $1 - \varepsilon^{1/2} \leq \theta \leq 1$, where a is a small constant. In the first two intervals we will prove that the derivative of $\Phi(\theta)$ with respect to θ is strictly negative. Recalling that $\|\varphi\|^2 \geq 1/2$, we have

$$\begin{aligned}
\frac{\mathrm{d}\Phi}{\mathrm{d}\theta} &\leq -\mathrm{atanh}\,\theta + k\alpha \mathbb{E}_\varphi \frac{(\varphi^{(1)}, \mathrm{T}_\theta\, \varphi^{(1)})}{\|\varphi\|^4} \\
&\leq -\theta + 2k\alpha \mathbb{E}_\varphi \frac{\sum_{i=1}^{k-1} |\varphi^{(1)}_{\{i\}}|^2}{\|\varphi\|^2}\, \theta + 2k\alpha \mathbb{E}_\varphi \frac{||\varphi^{(1)}||^2}{\|\varphi\|^2}\ \ \theta^3 \\
&\leq -\theta + Ae^{-Ck} \frac{\alpha}{\Omega_k}\theta + 2k\frac{\alpha}{\Omega_k}\theta^2 \leq -\frac{1}{2}\theta + 4k\theta^2 ,
\end{aligned}$$

where we used (from Eq. (2)) the hypothesis on low weight Fourier coefficients. The last expression is strictly negative if $0 < \theta < e^{-ak}$ for any $a > 0$ and all k large enough. The previous formula also shows $-\Phi\left(\theta\right) = \Omega\left(\theta^2\right)$ as $\theta \to 0$.

Next assume $e^{-ak} \leq \theta \leq 1 - \varepsilon$. Using the hypothesis $(\varphi^{(1)}, \mathrm{T}_\theta\, \varphi^{(1)}) \leq e^{-Ck(1-\theta)} ||\varphi^{(1)}||^2$, we have

$$\begin{aligned}
\frac{\mathrm{d}\Phi}{\mathrm{d}\theta} &\leq -\mathrm{atanh}\,\theta + 4k\alpha \mathbb{E}_\varphi \frac{||\varphi^{(1)}||^2}{||\varphi||^4}\ e^{-Ck\epsilon} \\
&\leq -\mathrm{atanh}\,\theta + 2k\frac{\alpha}{\Omega_k} e^{-Ck\sqrt{\epsilon}} \leq -\mathrm{atanh}\theta + 2\left(\log 2\right) k e^{-Ck\epsilon} ,
\end{aligned}$$

which is strictly negative if $\theta > e^{-ak}$ with, say, $a = (C\epsilon^2)/2$. Finally, we notice that, for $1 - \varepsilon^2 \leq \theta \leq 1$, any ε small enough we have $H(\theta) \leq -\log 2 + \varepsilon/10$. Further, using the fact that $(\varphi, \mathrm{T}_\theta\, \varphi) = ||\, \mathrm{T}_{\theta^{1/2}}\, \varphi||^2$ is non-decreasing in θ

$$\Phi(\theta) \leq -\log 2 + \frac{\varepsilon}{10} - \alpha \mathbb{E}_\varphi \log ||\varphi||^2 = -\log 2 + \frac{\varepsilon}{10} + \frac{\alpha}{\widehat{\Omega}_k} \leq -\varepsilon \frac{\log 2}{2} ,$$

which finishes the proof. □

Conclusion of Proof of Proposition 3.1. From the previous lemma we have that for any fixed $\delta > 0$,

$$\frac{\mathbf{E}\, Z_{\rm b}^2}{\left(\mathbf{E}\, Z_{\rm b}\right)^2} \leq \frac{\mathbf{E}\left[Z_{\rm b}(|Q_{12}| < (\delta/n)^{1/2})\right]}{\left(\mathbf{E}\, Z_{\rm b}\right)^2} + e^{-\Omega(\delta)},$$

while a calculation analogous to that in Eq. (22) and the fact that $-\Phi(\theta) = \Omega\left(\theta^2\right)$, implies that

$$\frac{\mathbf{E}\left[Z_{\mathrm{b}}(|Q_{12}| < (\delta/n)^{1/2})\right]}{\left(\mathbf{E}\, Z_{\mathrm{b}}\right)^2} \leq \sum_{\theta<(\delta/n)^{1/2}} \frac{C''}{n^{1/2}} \exp\left(-n\Omega\left(\theta^2\right)\right) \leq \int_{-\delta}^{\delta} \exp\left(-\Omega\left(x^2\right)\right) dx.$$

Now, letting $\delta \to 0$, it is clear that $\frac{\mathbf{E}\, Z_{\mathrm{b}}^2}{(\mathbf{E}\, Z_{\mathrm{b}})^2}$ tends to 1. This proves, by means of the Paley-Zygmund inequality, that for $\alpha < (1-\varepsilon)\left(\liminf_{n\to\infty} \Omega_{k,\widetilde{p}_n}\right) \log 2$, a $\mathrm{CSP}(n, p, \alpha; \widetilde{p}_n)$ ensemble is satisfiable w.h.p. The result extends straightforwardly for a random $\mathrm{CSP}(n, p, \alpha)$, after noticing that $\Omega_{k,L_n} > (1-\epsilon)\, \Omega_{k,p}$ with high probability. $\square$