Unsatisfiability Bounds for Random CSPs from an Energetic Interpolation Method

Dimitris Achlioptas[1,2,3,⋆] and Ricardo Menchaca-Mendez[3]

[1] University of Athens, Greece
[2] CTI, Greece
[3] University of California, Santa Cruz, USA

Abstract. The interpolation method, originally developed in statistical physics, transforms distributions of random CSPs to distributions of much simpler problems while bounding the change in a number of associated statistical quantities along the transformation path. After a number of further mathematical developments, it is now known that, in principle, the method can yield rigorous unsatisfiability bounds if one "plugs in an appropriate functional distribution". A drawback of the method is that identifying appropriate distributions and plugging them in leads to major analytical challenges as the distributions required are, in fact, infinite dimensional objects. We develop a variant of the interpolation method for random CSPs on arbitrary sparse degree distributions which trades accuracy for tractability. In particular, our bounds only require the solution of a 1-dimensional optimization problem (which typically turns out to be very easy) and as such can be used to compute explicit rigorous unsatisfiability bounds.

1 Introduction

The problem of determining the satisfiability of Boolean formulas is central to the understanding of computational complexity. Moreover, it is of tremendous practical interest as it arises naturally in numerous settings. Random CNF formulas have emerged as a mathematically tractable vehicle for studying the performance of satisfiability algorithms and proof systems. For a given set of n Boolean variables, let B_k denote the set of all possible disjunctions of k distinct, non-complementary literals from its variables (k-clauses). A random k-SAT formula $F_k(n, m)$ is formed by selecting uniformly and independently m clauses from B_k and taking their conjunction. Such random formulas have been shown to be hard both for proof systems, e.g., in the seminal work of Chvátal-Szemérdi on resolution [7], and, more recently, for some of the most sophisticated satisfiability algorithms known [8].

More generally, in Random Constraint Satisfaction Problems (RCSPs) one has a set of n variables all with the same (small) domain D and a set of $m = rn$ constraints, for some constant $r > 0$, each of which binds a randomly selected subset of $O(1)$ variables. Canonical examples are finding large independent sets and colorings sparse random graphs, variations of satisfiability, and systems of random linear equations. We will be interested in random CSPs (RCSPs) from an asymptotic point of view, i.e., as the

⋆ Research supported by NSF CCF-0546900, a Sloan Fellowship, and ERC grant 210743.

number of variables grows. In particular, we will say that a sequence of random events $\mathcal{E}_n$ occurs *with high probability (w.h.p.)* if $\lim \Pr[\mathcal{E}_n] = 1$. The ratio of constraints-to-variables, $r = m/n$, known as density, plays a fundamental role here as most interesting monotone properties are believed to exhibit 0-1 laws with respect to density. Perhaps the best known example is the satisfiability property for random k-CNF formulas. Let $g_k(n, r)$ denote the probability that $F_k(n, rn)$ is satisfiable.

Conjecture 1 (Satisfiability Threshold Conjecture). For each $k \geq 3$, there exists a constant r_k such that for any $\epsilon > 0$,

$$\lim_{n\to\infty} g_k(n, r_k - \epsilon) = 1, \quad \text{and} \quad \lim_{n\to\infty} g_k(n, r_k + \epsilon) = 0 \ .$$

The satisfiability threshold conjecture, which motivates our work, has attracted a lot of attention in computer science, mathematics and statistical physics [17,18,16]. At this point, neither the value, nor even the existence of r_k has been established. For $k = 3$, the best known bounds are $3.52 < r_3 < 4.49$, due to results in [9] and [13], respectively.

The last decade has seen a great deal of rigorous results on random CSPs, including a proliferation of upper and lower bounds for the satisfiability threshold of a number of problems. Equally importantly, random CSPs have been the domain of an extensive exchange of ideas between computer science and statistical physics [15], including the positing of the clustering phenomenon, establishing it rigorously, and relating it to algorithmic performance. In this work we take another step in this direction by taking a technique from mathematical physics, the *interpolation method* of Guerra [12], and using it to show how to derive end-to-end rigorous explicit upper bounds for the satisfiability threshold of a number of problems. To do so, we introduce a new version of the interpolation method that can be made computationally effective and give a new, much simpler, extension of the method to CSPs with arbitrary degree distributions.

Our method can be used to prove among other things the following result [5] regarding the satisfiability of mixtures of 2- and 3-clauses.

Theorem 1 ([5]). *Let F be a random CNF formula on n variables with $(1-\epsilon)n$ random 2-clauses, and $(1+\epsilon)n$ random 3-clauses. W.h.p. F is unsatisfiable for $\epsilon = 10^{-4}$.*

Theorem 1, combined with the methods of [3], implies that a number of DPLL algorithms require exponential time on easily satisfiable random 3-CNF formulas. For example, ORDERED DLL requires exponential time for all $r \geq 2.78$, while GUC for all $r \geq 3.1$. Similar results hold for a host of other algorithms, including, for example, all algorithms analyzed in [2] and [1].

2 Motivation and Past Work

Perhaps the simplest possible upper bound on the satisfiability threshold comes from taking the union bound over all assignments $\sigma \in \{0,1\}^n$ of the probability they satisfy a random formula $F = F_k(n, rn)$. That is,

$$\Pr[F_k(n, rn) \text{ is satisfiable}] \leq \sum_{\sigma} \Pr[\sigma \text{ satisfies } F_k(n, rn)] = \left[2(1 - 2^{-k})^r\right]^n \to 0 \ ,$$

for all $r > r_k^*$, where $2(1-2^{-k})^{r_k^*} = 1$. For example, $r_3^* = 5.19..$, but a long series of increasingly sophisticated results has culminated with the bound $r_3 < 4.48...$ by Díaz et al. [9]. At the same time, statistical physics results by Mertens et al. [14] give evidence that $r_3 < 4.26...$.

2.1 Past Work on the Interpolation Method for Random CSPs

The *interpolation method* is a remarkable tool originally developed by Guerra and Toninelli [12] to deal with the Sherrignton Kirkpatrick model (SK) of statistical physics. Following their breakthrough, Franz and Leone [10], in a very important paper, applied the interpolation method to random k-SAT and random k-XOR-SAT to prove that certain expressions derived via the non-rigorous replica method of statistical physics for these problems can, in principle, be used to derive upper bounds for the satisfiability threshold of each problem. As we will see, though, doing so involves the solution of certain functional equations that appear beyond analytical penetration. In [20], Panchenko and Talagrand showed that the results of [10] can be derived in a simpler and uniform way, unifying the treatment of different levels of Parisi's Replica Symmetry Breaking.

A crucial ingredient in all the above proofs is a Poissonization device exploiting that in Erdős-Renyi (hyper)graphs the degrees of the vertices behave, essentially, like independent, Poisson random variables. Franz, Leone, and Tonnineli [11] extended the interpolation method to other degree sequences, but at the cost of introducing another level of complexity (multi-overlaps), thus placing the method even further out of reach in terms of explicit computations. In [19], Montanari gave a simpler method for dealing with degree sequences in the context of error-correcting codes, which proceeds by approximating the intended degree distribution "in chunks". This, unfortunately, requires the number of approximation steps to go to infinity (so that the chunk size goes to zero) in order to give results for the original problem.

Finally, in a recent paper, Bayati, Gamarnik and Tetali [6], showed that a combinatorial analogue of the interpolation method can be used to elegantly derive an approximate subadditivity property for a number of CSPs on Erdős-Renyi and regular random graphs. This allowed them to prove the *existence* of a number of limits in these problems. The simplicity of that approach, though, comes at the cost of losing the capacity to give bounds for the associated limiting quantities.

3 Highlights of the Interpolation Method on RCSPs

For simplicity of exposition we focus on the case where all constraints have the same arity $k \geq 2$. It is very easy to see that the proof goes through transparently for CSPs that are mixtures of constraints of different arities. Let $C_{k,n}$ denote the set of all possible k-constraints on n variables for the CSP at hand and let D denote the domain of each variable. So, for example, $C_{k,n}$ could contain all $2^k\binom{n}{k}$ clauses of length k on n variables, or all $\binom{n}{2}D!$ possible unique-games constraints on a graph with n vertices. A random CSP instance $I_k(n,r)$ is a conjunction of m constrains taken independently with replacement from the set $C_{k,n}$, where m is a *Poisson random variable* with mean $\mathbb{E}[m] = rn$. Note that in the more standard models of random CSPs m is fixed (not a

random variable). Since, though, the standard deviation of the Poisson distribution is the square root of its mean we have $m = (1+o(1))rn$ w.h.p., thus not affecting any asymptotic results regarding densities. At the same time, along with the Poissonization of the variable degrees, this is key to the original development of the method. Eliminating the need for Poisson variable degrees and allowing arbitrary (sparse) degree sequence, as we do in Section 5, is part of the technical contribution in our work.

We shall work with the random variable $H_{n,r}(\sigma)$, known as the Hamiltonian, counting the number of unsatisfied constraints in the instance for each $\sigma \in D^n$. (The randomness of H being in the random choice of the instance). We will sometimes refer to $H_{n,r}(\sigma)$ as the energy function. The goal is to compute lower bounds for the following quantity as a function of $\beta \geq 0$,

$$f_r = f_r(\beta) = n^{-1}\mathbb{E}\left[\log\left(\sum_{\sigma\in D^n} \exp\left(-\beta H_{n,r}(\sigma)\right)\right)\right] \quad . \tag{1}$$

For each fixed value of $\beta > 0$, the sum in $f_r(\beta)$ is dominated by those value assignment having energy (violated constraints) in some narrow window that depends on β. (The idea being that assignments violating more constraints are penalized too heavily to contribute significantly to the sum, while assignment violating even fewer constraints are too rare to have substantial contribution.) Thus, $f_r(\beta)$ effectively counts the number of assignments at each energy level is known as the *free entropy density*. Note that as β is increased $f_r(\beta)$ places more and more weight to assignments violating fewer constraints, recovering the number of solutions as $\beta \to \infty$ (writing $\beta = 1/T$ this is also known as the zero-temperature limit). Standard martingale arguments imply that if any finite $\beta > 0$ we have $\lim_{n\to\infty} f_r(\beta) < 0$, then w.h.p. no solutions exist. The goal of the interpolation method is to give negative upper bounds for $f_r(\beta)$ and since f_r is the free entropy, we refer to this as entropic interpolation.

Given $\sigma = (x_1, x_2, \ldots, x_n)$ we will write $H_{n,r}(\sigma)$ as the sum of m functions $\theta_a(x_{a_1}, ..., x_{a_k})$, one for each constraint. That is, $\theta_a(x_{a_1}, ..., x_{a_k}) = 1$ if the associated constraint is not satisfied and 0 otherwise. For example, for k-SAT, we take the domain of the variables to be $\{+1, -1\}^n$ and for each k-clause $c_a(x_{a_1}, ..., x_{a_k})$ we let

$$\theta_a(x_{a_1}, ..., x_{a_k}) = \prod_{j=1}^{k} \frac{1 + J_{a_j} x_{a_j}}{2} \quad , \tag{2}$$

where $J_{a_{,j}} \in \{+1, -1\}$ represents the sign of literal a_j in clause c_a: $+1$ if the literal is negated and -1 otherwise.

The basic object of the interpolation method is a modified energy function that interpolates between $H_{n,r}(\sigma)$ and the energy function of a dramatically simpler and fully tractable model. Specifically, for $t \in [0, 1]$, let

$$\beta H_{n,r,t}(x_1, \ldots, x_n) = \sum_{m=1}^{m_t} \beta\theta_{a_m}(x_{a_{m,1}}, ..., x_{a_{m,k}}) + \sum_{i=1}^{n}\sum_{j=1}^{k_{i,t}} \log\left(\hat{v}_{i,j}(x_i)\right) \quad , \tag{3}$$

where m_t is a Poisson random variable with mean $\mathbb{E}[m_t] = trn$, the $k_{i,t}$'s are i.i.d. Poisson random variables with mean $\mathbb{E}[k_{i,t}] = (1-t)kr$, and the functions $\hat{v}_{i,j}(\cdot)$ are i.i.d. random functions distributed as the function defined in (5) below.

Before delving into the meaning of the random functions $\hat{v}_{i,j}(\cdot)$, which are the heart of the method, let us first make a few observations about (3). First, note that (3) is simply the energy function of the original model when $t = 1$. On the other hand, when $t < 1$, we expect that $(1 - t)m$ of the original k-clauses will be replaced by k times as many functions each of which takes as input a single variable. A helpful way to think about this replacement is as a decombinatorialization of the energy function wherein k-ary functions are replaced by univariate, and therefore, independent functions. As one can imagine, for $t = 0$ the model is fully tractable. In particular, letting

$$f_r(t) = n^{-1}\mathbb{E}\left[\log\left(\sum_{\sigma\in D^n}\exp\left(-\beta H_{n,r,t}(\sigma)\right)\right)\right] \ , \tag{4}$$

one can readily compute $f_r(0)$ since one can compute $H_{n,r,0}(\sigma)$ by examining one variable at a time. To relate the two models the plan is to give a lower bound for the change in f_r as t goes from 1 to 0, hence the name interpolation, thus bounding $f_r(1)$ by $f_r(0)$ plus a term depending on our bound on the derivative.

The main idea of the interpolation method is to select the (still mysterious) univariate functions $\hat{v}_{i,j}(\cdot)$ independently, from a probability distribution that reflects aspects of the geometry of the underlying solution space. The more accurate the reflection, the better the bound. One, of course, needs to guess this geometry and here is where the insights from statistical physics are most valuable. A beautiful aspect of the interpolation method is that it projects all information about the geometry of the solution space into a single object, a distribution γ as defined below. With that in mind, we now define the random univariate functions, *but without specifying the all-important distribution* γ. This is because the method gives a valid bound for *any* γ, i.e., the choice of γ affects the quality but not the validity of the derived bound.

Let $v(x)$ denote the density function of a random variable over D, where the probabilities $p_1, \ldots, p_{|D|}$ are themselves chosen at random from a distribution γ with support on the unit $(|D| - 1)$-dimensional simplex. Let $\hat{v}(x)$ be a random univariate function defined as follows

$$\hat{v}(x) = \sum_{y_1,\ldots,y_{k-1}}\exp\left(-\beta\theta(y_1, ..., y_{k-1}, x)\right)\prod_{j=1}^{k-1} v_j(y_j) \ , \tag{5}$$

where $\theta(\cdot)$ is a random constraint-function and the functions $v_i(\cdot)$ are i.i.d. with the same distribution as $v(x)$.

To interpret the function in (5) it helps to think of its argument x as corresponding to a particular occurrence of a variable in a constraint c, e.g., a literal occurrence in a random k-clause. The idea is for (5) to simulate the biases that this particular occurrence of x "feels" from its presence in c. To do this we replace c with a brand new random constraint (appearing as θ in (5)) containing $k-1$ new variables $y_1, \ldots, y_{k-1}$ which are "private" to θ, i.e., which will occur in no other constraint in the interpolating energy function. To simulate the statistical joint behavior of the $k - 1$ original variables in c due to their participation in clauses other than c, we assume that since the underlying random hypergraph is sparse, these $k - 1$ new variables are independent in the absence

of θ, hence the product in (5). Finally, specifying the probability distribution γ governing the behavior of each ersatz variable is precisely what reflects our beliefs about the geometry of the space of solutions. Statistical physics considerations suggest candidate distributions as solutions to distributional equations.

To see how the geometry of the space of solutions enters the distribution γ, consider two dramatically different settings, precisely those separated by the so-called shattering (or clustering, or dynamical) transition. In one setting, the set of solutions has the property that if a solution is chosen uniformly at random, changing the value of any variable to any other value can be accommodated by changing, in expectation, the value of $O(1)$ other variables, i.e., by "local repair". In such a world, γ is a single density function over the $(|D|-1)$-dimensional simplex. In contrast, after shattering occurs [4] the set of solutions consists of exponentially many clusters (connected components of solutions), separated by linear Hamming distance. In each cluster, a constant fraction of all variables take the same value in all solutions in the cluster, while all other variables are locally repairable. In this world, γ becomes a distribution over densities, the different densities corresponding to different clusters.

3.1 Why an Energetic Interpolation Method

What motivates our derivation of a different, so-called energetic, interpolation method is that dealing with the shattered case above leads to massive analytical obstacles, rendering the derivation of explicit, mathematically rigorous bounds problematic. In the realm of statistical mechanics, these are addressed via a numerical stochastic method known as population dynamics, used to derive the estimates in [14] for random k-SAT.

In contrast, we will see that the energetic approach leads to bounds which can be derived analytically, precisely because we dramatically collapse the information captured by γ. In particular, in our bounds γ will be specified by a single real number, while the bound itself is expressed by truncating an infinite sum to a finite one (at any desired degree of accuracy) and adding up the corresponding explicit terms, each involving the joint behavior of a finite number of Poisson random variables.

The reason this approach works is that in bivariate binary CSPs, such as random [MAX] 2-SAT, random MAX 2-LIN-2, and random $(2+p)$-SAT, whenever a frozen variable appears in a constraint "the wrong way" (the freezing being due to its participation in other constraints) this necessarily causes the other variable in the constraint to also freeze. This percolative type of behavior causes the fraction of frozen variables to take off smoothly in such problems, a situation that can be captured by a simple model for the distribution γ if one focuses on states of lowest energy. This is precisely what we exploit in deriving our new upper bounds for these problems.

4 Energetic Interpolation for General CSPs

To develop an energetic interpolation method we replace the (far richer) free entropy density of the previous section with the following much simpler quantity

$$\xi_r = n^{-1}\mathbb{E}\left[\min_{\sigma\in D^n} H_{n,r}(\sigma)\right] \quad , \tag{6}$$

known as ground-state energy density, which simply tells us the fraction of violated constraints in the optimal (least-violating) assignments. By standard martingale arguments the random variable $\min_\sigma H_{n,r}(\sigma)$ concentrates around its expectation (consider the martingale exposing the constraints one by one and note that changing any one constraint cannot change its value by more than 1). Therefore, if $\liminf_{n\to\infty} \xi_r > 0$ we can conclude that the satisfiability threshold is upper bounded by r.

The univariate factors in the energy interpolation method are given as follows:

- For $1 \leq j \leq |D|$, let "j" denote the indicator function that the input is j, i.e., "j" is 1 if its input is j and 0 otherwise.
- Let "*" denote the function that assigns 0 to all elements of D.
- Let $\mathtt{h}(x)$ be a random function in $\{\text{“1”}, \ldots, \text{“}|D|\text{”}, \text{“}*\text{”}\}$ with $\Pr(\mathtt{h}(\cdot) = \text{“}*\text{”}) = 1 - p$ and $\Pr(\mathtt{h}(\cdot) = \text{“}j\text{”}) = p/|D|$.

The analogue of (5) is now

$$\hat{h}(x) = \min_{y_1,\ldots,y_{k-1}} \left\{ \theta(y_1, .., y_{k-1}, x) + \sum_{i=1}^{k-1} \mathtt{h}_i(y_i) \right\} \quad , \tag{7}$$

where $\theta(\cdot)$ is a random constraint-function as before while the functions $\mathtt{h}_i(\cdot)$ are i.i.d. random functions distributed as $\mathtt{h}(x)$.

Observe that the energy interpolation method models all information about the geometry of the solution space into a single probability p, which can be interpreted as the probability that a variable picked at random will be frozen, i.e., have the same value in all optimal assignments. If that occurs for all $k-1$ variables $y_1, .., y_{k-1}$ and they all happen to be frozen the wrong way as far as θ is concerned, then unless variable x takes the value desired by θ the function $\hat{h}(x)$ will evaluate to 1. When, at the end of the interpolation, we will have replaced all k-ary constraints with univariate random functions $\hat{h}$, the optimal overall assignment is simply found by assigning to each variable the value that makes the majority of its $\hat{h}$ functions evaluate to 0. The method delivers a valid bound for *any* choice of $p \in [0, 1]$ and the bound is then optimized by choosing the best value of p, i.e., performing a single-parameter search.

While we could give lower bounds on (6) for RCSPs defined on Erdős-Renyi (hyper)graphs by exploiting the same Poissonization device as in earlier works, we will instead show how to carry out the method in arbitrary sparse degree distributions.

5 The Interpolation Method on Sparse Degree Sequences

Let d_i denote the number of times variable i should appear in the random instance and let $L_i = \{l_{i,j}\}_{j=1}^{d_i}$ denote the set of occurrences corresponding to variable i. Note that the occurrences can be decorated so that, for example in k-SAT, we can specify how many of the L_i occurrences correspond to positive occurrences of the variables and how many to negative occurrences. It will be helpful to think of each occurrence as a piece of paper carrying the index of the underlying variable along with any desired decoration. To form a random instance with $m = rn$ constraints we simply choose a

random permutation of the krn elements of $\mathcal{L} = \{L_i\}_{i=1}^{n}$ and consider the first k to specify the first constraint, the next k to specify the second constraint etc.

Consider now the following algorithm to build a random Hamiltonian composed of a mixture of k-ary constraint-factors of the desired CSP and of univariate functions as in (5). The algorithm has three inputs: The collection of occurrences $\mathcal{L}$, an integer t, and a sequence $x \in \{b, c\}^t$.

1. Set $H = \emptyset$, set $L = \mathcal{L}$, and set $j = 1$.
2. Select a random permutation π of the elements of L.
3. **While** $j \leq \min\{t, |L|\}$ **do**:
   (a) **If** $x_j = b$ **then**
      i. Add a random univariate factor to H with argument $\pi(j)$.
      ii. $j \leftarrow j + 1$
   (b) **If** $x_t = c$ **then** with probability $1/k$
      i. Add a random k-constraint to H on occurrences $\pi(j), \ldots, \pi(j + k - 1)$.
      ii. $j \leftarrow j + k$

Let $\mathbb{H}(\mathcal{L}, x)$ denote the family of energy functions produced by the above algorithm. Observe that when $t = |\mathcal{L}|$ and $x = u \cdots u$, the energy functions produced by the algorithm have variable degree distribution given by $\mathcal{L}$ and consist of univariate factors only. On the other hand when $t = |\mathcal{L}|$ and $x = c \cdots c$ the resulting energy functions consist of $\tilde{m}$ energy constraint functions of arity k where $\tilde{m}$ is a Binomial random variable with km trials and probability of success $1/k$, conditioned on being at most m. In other words, w.h.p. the instance generated will have the desired degree sequence except for $o(n)$ variables (and, therefore, $o(n)$ constraints). Since we are interested in establishing a non-vanishing lower bound for (6) this will not affect any of our results.

The goal now is to relate the ground state energy of these two extreme cases. A key property, which will allow us to establish such relation, is that $\mathbb{H}(\mathcal{L}, x)$ is invariant under any permutation $\pi(x)$ of the elements in x.

Lemma 1. *For every sequence x, and every permutation π, the families $\mathbb{H}(\mathcal{L}, x)$ and $\mathbb{H}(\mathcal{L}, \pi(x))$ have the same distribution.*

Proof. The very first step of our construction is to take a uniformly random permutation of the elements of $\mathcal{L}$.

For any $\mathcal{L}$ and any $s \leq t$, since the order of the steps in x does not matter, let us write $\mathbb{H}(\mathcal{L}, t, s)$ to denote the distribution of energy functions generated by the algorithm when we take t steps in total, $t - s$ of which are additions of a univariate factor. Let

$$\xi_{\mathcal{L}}(t, s) = n^{-1}\mathbb{E}\left[\min_{\sigma \in D^n} H_{\mathcal{L},t,s}(\sigma)\right] \quad .$$

Observe that if $t = km$ and $s = km$, then $\xi_{\mathcal{L}} = \xi_{\mathcal{L}}(km, km)$ corresponds to the original ground state energy, whereas $\xi_{\mathcal{L}}(km, 0)$ corresponds to the ground state energy of the model composed of univariate factors only.

Our lower bounds come from the following theorem.

Theorem 2. *For any choice of $p \in [0,1]$, if $m = rn$ then*

$$\xi_{\mathcal{L}} \geq \xi_r(km, 0) - r(k-1)\mathbb{E}\left[h_c\right] - o(1) \ , \tag{8}$$

where

$$h_c = \min_{y_1,\ldots,y_k} \left\{ \theta(y_1, .., y_k) + \sum_{i=1}^{k} \mathtt{h}_i(y_i) \right\} \ .$$

To prove this we will prove that as s goes from t to 0, we can control the change of $\xi_r(t, s)$. Specifically,

Lemma 2. *If $m = rn$ then for any $\epsilon > 0$, all $t \in [0, (kr - \epsilon)n]$, and all $1 \leq s \leq t$,*

$$\mathbb{E}\left[\min\{H_{\mathcal{L},t,s-1}(\sigma)\}\right] - (k-1)k^{-1}\mathbb{E}\left[h_c\right] \leq \mathbb{E}\left[\min\{H_{\mathcal{L},t,s}(\sigma)\}\right] + o(1) \ .$$

Iteratively applying Lemma 2 so that we can increase the number of univariate factors from 0 to $t = (kr - \epsilon)n$ and letting $\epsilon \to 0$ yields Theorem 2.

Proof (Lemma 2). Let H_0 be an energy function from the family $\mathbb{H}(\mathcal{L}, t-1, s-1)$, that is, the energy function resulting from executing $t-1$ steps of the algorithm where $s-1$ of such steps correspond to adding a univariate factor. The key observation is that $H_{\mathcal{L},t,s-1}(\sigma)$ and $H_{\mathcal{L},t,s}(\sigma)$ can be obtained from H_0 by execution an additional step of the algorithm: $H_{\mathcal{L},t,s-1}(\sigma)$ corresponds to the processing of a c symbol and $H_{\mathcal{L},t,s}(\sigma)$ corresponds to the precessing of a u symbol.

We will show that conditional on any realization of H_0 we have

$$\mathbb{E}\left[\min\{H_{\mathcal{L},t,s-1}(\sigma)\}|H_0\right] - (k-1)k^{-1}\mathbb{E}\left[h_c\right] \leq \mathbb{E}\left[\min\{H_{\mathcal{L},t,s}(\sigma)\}|H_0\right] + o(1) \ . \tag{9}$$

That is, the proof reduces to comparing the effect of adding a single univariate factor to the effect of adding, with probability $1/k$, a single constraint. As one can imagine, the proof of (9) is problem specific. Below we prove it for random k-SAT and random Max-k-Lin-2. For all other random CSPs with binary domains the proof is very similar.

6 Applying Energetic Interpolation to Random CSPs

6.1 Random *k*-SAT

Let $C^* \subseteq \{0,1\}^n$ be the set of optimal assignments in H_0. A variable x_i is frozen if its value is the same in all optimal assignments. The processing of a c symbol will increase the value of the minimum by at most 1 only if the following two conditions hold: 1) a new clause is added, which occurs with probability $1/k$, and 2) all the literals appearing in the new random factor correspond to a frozen variables. By the principle of deferred decisions we can think of the permutation π as generated on-the-fly, i.e., as we need occurrences to consume. Therefore, if the number of remaining occurrences is $\Omega(n)$ and f denotes the fraction of them that are associated with frozen variables corresponding to H_0, then

$$\mathbb{E}\left[\min\{H_{\mathcal{L},t,s}(\sigma)\}|H_0\right] - \min\{H_0\} = k^{-1}2^{-k}f^k + O(1/n) \ ,$$

where the last term is due to the fact that we are selecting without replacement.

Similarly, the processing of a u symbol will increase the value of the minimum by 1 if the chosen literal correspond to a frozen variable x and x must take the opposite of its frozen value to minimize the added factor $\hat{h}(x)$. Thus the expected change is

$$\mathbb{E}\left[\min\{H_{\mathcal{L},t,s-1}(\sigma)\}|H_0\right] - \min\{H_0\} = 2^{-k}p^{k-1}f \ .$$

Finally,

$$\mathbb{E}\left[h_c\right] = \mathbb{E}\left[\min_{y_1,\ldots,y_k}\{\theta(y_1,..,y_k) + \sum_{i=1}^{k}\mathtt{h}_i(y_i)\}\right] = 2^{-k}p^k \ .$$

By combining the above equations and adding $-(k-1)k^{-1}2^{-k}p^k$ we get

$$\begin{aligned}\mathbb{E}\left[\min\{H_{\mathcal{L},t,s-1}(\sigma)\}|H_0\right] - (k-1)k^{-1}2^{-k}p^k - \mathbb{E}\left[\min\{H_{\mathcal{L},t,s}(\sigma)\}|H_0\right] \\ = k^{-1}2^{-k}\left(kp^{k-1}f - f^k - (k-1)p^k\right) + O(1/n) \ .\end{aligned}$$

Finally, the polynomial $F(x,p) = kp^{k-1}x + x^k - (k-1)p^k \leq 0$ for all $0 \leq x, p \leq 1$. To see this last statement note that (i) $F(0,p), F(1,p), F(x,0), F(x,1) \leq 0$ and, (ii) the derivative of F with respect to p is 0 only when $p = x$, in which case $F(x,x) = 0$.

6.2 Random Max-k-Lin-2

The constraints in the Max-k-Lin-2 problem are chosen uniformly from the set of all $2n^k$ possible boolean equations on n variables, i.e., the k variables are chosen at random with replacement and the required parity is equally likely to be 0 or 1. Let $C^* \subseteq \{0,1\}^n$ be the set of optimal assignments in H_0. A variable x_i is frozen if its value is the same in all optimal assignments. The processing of a c symbol will increase the value of the minimum by at most 1 only if the following three conditions hold: 1) a new Boolean equation is added, which occurs with probability $1/k$, 2) all the literals appearing in the new random factor correspond to frozen variables and 3) the parity of the frozen variables is different from the one required by the new equation. As in the proof for random k-SAT above, if the number of remaining occurrences is $\Omega(n)$ and f denotes the fraction of them that are associated with frozen variables corresponding to H_0, then,

$$\mathbb{E}\left[\min\{H_{\mathcal{L},t,s}(\sigma)\}|H_0\right] - \min\{H_0\} = k^{-1}2^{-1}f^k + O(1/n) \ .$$

where the last term is due to the fact that we are selecting without replacement. Similarly, the processing of a c symbol can increase the value of the minimum by 1 if the chosen literal correspond to a frozen variable x and and x must take the opposite of its frozen value to minimize the added factor $\hat{h}(x)$. Thus the expected change is given by

$$\mathbb{E}\left[\min\{H_{\mathcal{L},t,s-1}(\sigma)\}|H_0\right] - \min\{H_0\} = 2^{-1}p^{k-1}f \ .$$

Finally,

$$\mathbb{E}\left[h_c\right] = \mathbb{E}\left[\min_{y_1,\ldots,y_k}\{\theta(y_1,..,y_k) + \sum_{i=1}^{k}\mathtt{h}_i(y_i)\}\right] = 2^{-1}p^k \ .$$

Combining the above equations and adding $-(k-1)k^{-1}2^{-1}p^k$ we get

$$\begin{aligned}\mathbb{E}\left[\min\{H_{\mathcal{L},t,s-1}(\sigma)\}|H_0\right] - (k-1)k^{-1}2^{-k}p^k - \mathbb{E}\left[\min\{H_{\mathcal{L},t,s}(\sigma)\}|H_0\right] \\ = k^{-1}2^{-1}\left(kp^{k-1}f - f^k - (k-1)p^k\right) + O(1/n) \ ,\end{aligned}$$

where the r.h.s. of the equality entails the same polynomial as for random k-SAT.

7 Computing Explicit Energetic Interpolation Bounds for k-SAT

Applying Theorem 2 on a Poisson degree sequence we get that

$$\xi_r(0) = \mathbb{E}\left[\min_{x\in\{0,1\}}\left(\sum_{j=1}^{s}\hat{h}_j(x)\right)\right] \ , \tag{10}$$

where s is a Poisson random variable with mean kr, and the functions $\hat{h}_j(\cdot)$, i.e., random functions in $\{\text{“0”}, \text{“1”}, \text{“}*\text{”}\}$ with $\Pr(\hat{h}_j(\cdot) = \text{“1”}) = \Pr(\hat{h}_j(\cdot) = \text{“0”}) = 2^{-k}p^{k-1}$.

Let l_0, l_1, and l_* denote the number “0”, “1”, and “*” functions respectively among the $\hat{h}_j(\cdot)$ functions inside the summation of equation (10). Conditional on the value of s, the random vector (l_0, l_1, l_*) is distributed as a multinomial random vector and, therefore,

$$\xi_r(0) = \sum_{x=0}^{\infty}\sum_{l_0=0}^{x}\sum_{l_1=0}^{x-l_0}\min\{l_0,l_1\}\times\text{Poi}(kr,x)\text{Multi}(l_0,l_1,x-l_0-l_1) \ ,$$

where $\text{Multi}(\cdot,\cdot,\cdot)$ denotes the multinomial density function.

Changing the limits of all summations to infinity, does not change the value of $\xi_r(0)$, since $\text{Multi}(\cdot,\cdot,\cdot)$ evaluates to zero for negative numbers, hence, we can interchange the order of the summations to get

$$\xi_r(0) = \sum_{l_0=0}^{\infty}\sum_{l_1=0}^{\infty}\min\{l_0,l_1\}\times\sum_{x=0}^{\infty}\text{Poi}(kr,x)\text{Multi}(l_0,l_1,x-l_0-l_1) \ .$$

The last equation can be simplified by summing out the randomness in the Poisson random variable. The result is that l_0 and l_1 become two independent Poisson random variables with mean $\frac{k}{2^k}rp^{k-1}$. Thus,

$$\xi_r(0) = \sum_{l_0=0}^{\infty}\sum_{l_1=0}^{\infty}\min\{l_0,l_1\}\times\text{Poi}\left(\frac{k}{2^k}rp^{k-1},l_0\right)\times\text{Poi}\left(\frac{k}{2^k}rp^{k-1},l_1\right) \ ,$$

i.e., $\xi_r(0)$ is the expected value of the minimum of two independent Poisson random variables l_0, l_1 with mean $\lambda = \frac{k}{2^k}rp^{k-1}$. Finally, we note that

$$\mathbb{E}\left[\min\{l_0,l_1\}\right] = \sum_{i=0}^{\infty} i\left(2\text{Poi}(\lambda,i)\left(1-\sum_{j=0}^{i-1}\text{Poi}(\lambda,j)\right) - \left(\text{Poi}(\lambda,i)\right)^2\right) \ . \tag{11}$$

To compute a rigorous lower bound for (11) one now simply truncates the sum at any desired level of accuracy.

Acknowledgements. We are grateful to Andrea Montanari for a number of useful conversations.

References

1. Achioptas, D., Sorkin, G.: Optimal myopic algorithms for random 3-sat. In: Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on. pp. 590–600. IEEE (2000)
2. Achlioptas, D.: Lower bounds for random 3-sat via differential equations. Theoretical Computer Science 265(1-2), 159–185 (2001)
3. Achlioptas, D., Beame, P., Molloy, M.: A sharp threshold in proof complexity yields lower bounds for satisfiability search. Journal of Computer and System Sciences 68(2), 238–268 (2004)
4. Achlioptas, D., Coja-Oghlan, A.: Algorithmic barriers from phase transitions. In: 49th Annual IEEE Symp. on Foundations of Computer Science, 2008. pp. 793–802 (2008)
5. Achlioptas, D., Menchaca-Mendez, R.: Exponential lower bounds for dpll algorithms on satisfiable random 3-cnf formulas (2012), to appear in SAT12
6. Bayati, M., Gamarnik, D., Tetali, P.: Combinatorial approach to the interpolation method and scaling limits in sparse random graphs. In: STOC'10. pp. 105–114 (2010)
7. Chvatal, V., Szemeredi, E.: Many hard examples for resolution. Journal of the Association for Computing Machinery 35(4), 759–768 (1988)
8. Coja-Oghlan, A.: On belief propagation guided decimation for random k-sat. In: Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms. pp. 957–966. SIAM (2011)
9. Díaz, J., Kirousis, L., Mitsche, D., Pérez-Giménez, X.: On the satisfiability threshold of formulas with three literals per clause. Theoretical Computer Science 410(30-32), 2920–2934 (2009)
10. Franz, S., Leone, M.: Replica bounds for optimization problems and diluted spin systems. Journal of Statistical Physics 111(3), 535–564 (2003)
11. Franz, S., Leone, M., Toninelli, F.: Replica bounds for diluted non-poissonian spin systems. Journal of Physics A: Mathematical and General 36, 10967 (2003)
12. Guerra, F., Toninelli, F.: The thermodynamic limit in mean field spin glass models. Communications in Mathematical Physics 230(1), 71–79 (2002)
13. Kaporis, A., Kirousis, L., Lalas, E.: The probabilistic analysis of a greedy satisfiability algorithm. Random Structures & Algorithms 28(4), 444–480 (2006)
14. Mertens, S., Mézard, M., Zecchina, R.: Threshold values of random k-sat from the cavity method. Random Structures & Algorithms 28(3), 340–373 (2006)
15. Mezard, M., Montanari, A.: Information, physics, and computation. Oxford University Press, USA (2009)
16. Monasson, R., Zecchina, R.: Tricritical points in random combinatorics: the-sat case. Journal of Physics A: Mathematical and General 31, 9209 (1998)
17. Monasson, R., Zecchina, R.: Entropy of the K -satisfiability problem. Phys. Rev. Lett. 76, 3881–3885 (May 1996), http://link.aps.org/doi/10.1103/PhysRevLett.76.3881
18. Monasson, R., Zecchina, R.: Statistical mechanics of the random k-satisfiability model. Phys. Rev. E 56, 1357–1370 (Aug 1997), http://link.aps.org/doi/10.1103/PhysRevE.56.1357
19. Montanari, A.: Tight bounds for ldpc and ldgm codes under map decoding. Information Theory, IEEE Transactions on 51(9), 3221 –3246 (sept 2005)
20. Panchenko, D., Talagrand, M.: Bounds for diluted mean-fields spin glass models. Probability Theory and Related Fields 130(3), 319–336 (2004)