Adversarial satisfiability problem

Michele Castellana\textsuperscript{1,2,3} and Lenka Zdeborová\textsuperscript{3,4}

\textsuperscript{1} Dipartimento di Fisica, Università di Roma ‘La Sapienza’, 00185 Rome, Italy
\textsuperscript{2} LPTMS, CNRS and Université Paris-Sud, UMR8626, Bâtiment 100, 91405 Orsay, France
\textsuperscript{3} Theoretical Division and Center for Nonlinear Studies, Los Alamos National Laboratory, NM 87545, USA
\textsuperscript{4} Institut de Physique Théorique, IPhT, CEA Saclay, and URA 2306, CNRS, 91191 Gif-sur-Yvette cedex, France

E-mail: michele.castellana@lptms.u-psud.fr and lenka.zdeborova@cea.fr

Received 4 November 2010
Accepted 26 February 2011
Published 28 March 2011

Online at stacks.iop.org/JSTAT/2011/P03023
doi:10.1088/1742-5468/2011/03/P03023

Abstract. We study the adversarial satisfiability problem, where the adversary can choose whether the variables are negated in clauses or not, in order to make the resulting formula unsatisfiable. This problem belongs to a general class of adversarial optimization problems that often arise in practice and are algorithmically much harder than the standard optimization problems. We use the cavity method to compute large deviations of the entropy in the random satisfiability problem with respect to the configurations of negations. We conclude that in the thermodynamic limit the best strategy the adversary can adopt is to simply balance the number of times every variable is negated and the number of times it is not negated. We also conduct a numerical study of the problem, and find that there are very strong pre-asymptotic effects that may be due to the fact that for small sizes exponential and factorial growth is hardly distinguishable. As a side result we compute the satisfiability threshold for balanced configurations of negations, and also the random regular satisfiability, i.e. when all variables belong to the same number of clauses.

Keywords: message-passing algorithms, random graphs, networks
1. Introduction

The following setting often arises in practical optimization problems. Consider two players, each of them has a given set of moves (configurations) and a cost function depending on the moves of both the players. The first player is trying to optimize a certain cost function over his set of moves (configurations), and the interest of the second player is to make this optimum as bad as possible. In the game, at first the second player (adversary) chooses his moves (a configuration), and then the first player chooses his moves. What is the best strategy (algorithm) for the adversary in the case where his set of moves is too large to be able to evaluate all the possibilities? A specific example of this adversarial optimization setting could be a police department trying to set border controls in such a way that the amount of goods smugglers can transfer is the smallest possible [1, 2], or minimax games treated in [3].

Let us call the set of moves of the adversary \( \vec{u} \), and the set of moves of the first player \( \vec{v} \), the cost function being \( f(\vec{u}, \vec{v}) \). The goal of the adversary is to find \( u_m \) that maximizes \( \min_\vec{v} f(\vec{u}, \vec{v}) \). Since in common situations both \( \vec{u} \) and \( \vec{v} \) have exponentially many components in the size of the system the adversarial optimization is much harder than the usual (one-player) optimization, because even evaluating the \( \min_\vec{v} f(\vec{u}, \vec{v}) \) for a given \( \vec{u} \) is typically a hard optimization problem.

In the theory of algorithmic complexity the so-called NP problems are those for which it is easy (polynomial) to evaluate if a proposed solution is indeed a solution. In other words verifying a solution to an NP problem is a polynomial problem. Assume now that for a given \( \vec{u} \) the problem \( \exists \vec{v} \Phi(\vec{u}, \vec{v}) = \text{TRUE} \) is an NP problem (as an example consider \( \Phi(\vec{u}, \vec{v}) = \text{TRUE} \) if and only if \( f(\vec{u}, \vec{v}) < a \) where \( f \) is the cost function
Adversarial satisfiability problem

from above, and \( a \) is a constant). Then the associated adversarial decision problem is defined as \( \exists \vec{u} \forall \vec{v} \Phi(\vec{u}, \vec{v}) = \text{TRUE} \). Hence if the class of polynomial problems is considered easy, and the NP problems correspond to the first level of difficulty, then the adversarial problems correspond to the second level of difficulty. We could continue this construction by adding even more levels of logical quantifiers; this would lead to the so-called polynomial hierarchy of complexity classes. For more details on these notions see for instance [4,5]. Without doubt, the theoretical understanding of hard optimization and decision problems is crucial for many areas of science, and the same holds for adversarial problems.

The most famous benchmark of an algorithmically hard problem is the \( K \)-satisfiability (\( K \)-SAT) of Boolean formulas. Call a clause a logical disjunction (operation ‘or’) of \( K \) variables or their negations. Given a set of \( N \) Boolean variables \( x_i \), and a set of \( M \) clauses, the satisfiability problem consists in deciding whether all clauses can be simultaneously satisfied or not. The \( K \)-SAT problem was the first problem shown to be NP-complete, that is as hard as any other NP problems [6]. It has a large number of applications in automated verification and design. The random \( K \)-SAT problem, where variables in clauses are chosen randomly and negated with probability \( 1/2 \), provides easy to generate hard formulas [7].

Random \( K \)-SAT thus became a common playground for new algorithms, and theoretical ideas for understanding the origin of algorithmic hardness. The statistical physics approach related to the physics of diluted spin glasses contributed tremendously to the understanding of the properties of random \( K \)-SAT formulas, see e.g. [8,9]. Following this path in this paper we introduce a random adversarial satisfiability problem and develop a statistical mechanics framework to understand its properties. This framework can be readily applied to other random adversarial optimization and decision problems.

We study the large deviation functions for the original optimization problem with respect to the moves of the adversary. The main ideas of our approach to the study of large deviations come from studies of spin glasses and the cavity method [10]–[13]. Our approach is also closely linked to the well established fact that the replicated free energy in Parisi’s replica symmetry breaking (RSB) [14,15] can be interpreted as Legendre transformation of the large deviation function. We will, however, derive our method independently of these notions, by using only the factor graph representation of the problem and the belief propagation (BP) algorithm.

One natural setting for the random adversarial satisfiability problem is to introduce the negation-variables \( J_{ia} \), where \( J_{ia} = 1 \) if variable \( i \) is negated in clause \( a \), and \( J_{ia} = 0 \) if not. The set of moves of the adversary is then all the possible configurations of negations \( \{J_{ia}\} \), while the moves of the first player are all the possible configurations of the variables \( \{x_i\} \). The graph of interactions is chosen at random as before. The goal of the adversary is to set the negations in such a way that the resulting formula is as frustrated as possible. In particular, we will be interested in the question: can the adversary make the formula unsatisfiable or not? We will call this problem random AdSAT.

An independent interest in the random AdSAT comes from the study of the random quantum satisfiability (quantum SAT) problem [16]–[18]. It was shown that if the adversary can make the random formula unsatisfiable, then also the random quantum SAT is unsatisfiable. A natural question is whether the quantum SAT is much more restrictive than the AdSAT or not.

doi:10.1088/1742-5468/2011/03/P03023
Adversarial satisfiability problem

Note also that the quantified satisfiability (QSAT) problem is another SAT-based problem that belongs to the general adversarial setting as introduced above. In the QSAT problem one introduces two types of variables: the existential variables $x_i$ and the universal variables $y_i$. The QSAT then consists in deciding whether $\forall \vec{y} \exists \vec{x} G(\vec{x}, \vec{y}) = \text{TRUE}$, where $G(\vec{x}, \vec{y})$ is a satisfiability formula (with negation-variables fixed). Random ensembles of QSAT were introduced and studied in [19]. QSAT is arguably more important for industrial applications than the AdSAT that we study here. We chose the AdSAT defined as above because it is slightly simpler, and it provides information relevant to the original random SAT problem. We plan to apply our approach to the random QSAT problem in the near future.

In terms of methodology our study is also closely related to the work on optimization under uncertainty [20], where the goal is to find $\arg\min_{\vec{u}} \mathbb{E}_t \min_{\vec{v}} f(\vec{u}, \vec{v}, t)$. In other words one needs to minimize an expectation of a result of an independent minimization, which itself can be hard to compute. In [20] the authors studied the stochastic bipartite matching problem with a message passing technique very closely related to the one we use for the adversarial SAT problem here.

The present paper is structured as follows. In section 2 we set the adversarial satisfiability problem, and describe our statistical physics approach to solve it. In section 3 we recall the standard belief and survey propagation (SP) equations for the random $K$-satisfiability problem. In section 4 we derive the equations to compute the large deviations with respect to the configurations of negations. In section 5 we first present and discuss the cavity result for random regular adversarial satisfiability, then we do the same for the canonical (Poissonian) random adversarial satisfiability. As a side result we obtain the satisfiability threshold for the (non-adversarial) random regular and random balanced SAT. In section 6 we compare our theoretical result to numerical simulations, performing an exhaustive search of all the solutions. Finally, in section 7 we conclude and discuss the perspectives of this work.

2. AdSAT as a large deviation calculation

The random $K$-SAT problem is defined as follows. Consider $N$ Boolean variables $\{x_i\}_{i=1,\ldots,N}$, $x_i = \{0,1\}$, and $M = \alpha N$ clauses $\psi_a$. Each clause depends on $K$ random variables from the $N$ available ones. If a variable $i$ belongs to clause $\psi_a$, then we set $J_{ai} = 1$ if the variable is negated, and $J_{ia} = 0$ if it is not. The $K$-SAT problem can be represented via a so-called factor graph, a bipartite graph between variables (variable nodes) and clauses (function nodes), with edges between variable $i$ and clause $a$ if $i$ belongs to clause $a$. The negation-variables can be seen as attributes of the edges. The random $K$-SAT instance corresponds to the case where the $J_{ia}$s are drawn uniformly at random. Probably, the most well known property of the random $K$-SAT is the existence of a phase transition at a value $\alpha_c$ such that if $\alpha < \alpha_c$ then with high probability (probability going to one as $N \to \infty$) there exists a configuration $\{x_i\}$ that satisfies all the clauses, and for $\alpha > \alpha_c$ no such configuration exists with high probability. We define $\partial_i$ as the ensemble of function nodes connected to the variable node $i$, $\partial_a$ as the ensemble of variable nodes connected to the function node $a$.

The adversarial satisfiability problem (AdSAT) is defined by drawing a random $K$-SAT instance as before without deciding the negation-variables $\{J_{ia}\} \equiv J$. A solution to
Adversarial satisfiability problem

the AdSAT problem is given by a set $\mathcal{J}$ such that the resulting instance is unsatisfiable. Just as in random $K$-SAT there is a threshold $\alpha_a$ in the random AdSAT such that for $\alpha < \alpha_a$ no solution to the AdSAT formula exists with high probability. And for $\alpha > \alpha_a$ a solution exists with high probability. We observe that $\alpha_a \leq \alpha_c$ since above $\alpha_c$ a random configuration of negations makes the formula unsatisfiable, recall $\alpha_c(K = 3) = 4.2667$ [21]. Also $\alpha_a \geq \alpha_p$, where $\alpha_p = 1/K$ is the percolation threshold below which the graph is basically a collection of small trees and a few single loop components, which are both satisfiable for any configuration of negations. One of the goals of the present paper is to estimate the value of the AdSAT threshold $\alpha_a$.

In random $K$-SAT the satisfiability threshold can be found by counting the number of configurations that have a certain energy $E(\{x_i\})$ (i.e. number of unsatisfied clauses). To compute the entropy one introduces a Legendre parameter $\beta$ and computes the free energy $f$ defined as

$$e^{-\beta N f(\beta)} = \sum_{\{x_i\}} e^{-\beta E(\{x_i\})} = e^{N[s(e) - \beta e]}, \quad \frac{\partial s(e)}{\partial e} = \beta, \quad (1)$$

where the number of configurations having energy $E$ is $e^{s(E)}$, and the saddle-point approximation for $N \to \infty$ has been used. If $E = 0$ belongs to the support of the function $S(E)$ then the problem is in the satisfiable phase, if not than the problem is in the unsatisfiable phase. In the satisfiable phase we call $s = S(0)/N$ the entropy of satisfying configurations. The cavity method and the replica symmetry breaking serve to compute $f(\beta)$ in the thermodynamic limit $N \to \infty$ [22,8]. There are two crucial properties that make this calculation possible. First, the energy can be written as a sum of local terms:

$$E(\{x_i\}) = \sum_a \prod_{i \in \partial a} \delta_{x_i, J_{ia}}. \quad (2)$$

Second, the underlying factor graph is locally tree-like. These computations moreover provide much more information about the problem than the value of the satisfiability threshold.

In the study of random AdSAT we will proceed analogously. We consider the number of configurations of the negations that yield a given value of the entropy of solutions $s$

$$s(\mathcal{J}) = \frac{1}{N} \log \left[ \sum_{\{x_i\}} \prod_{a=1}^{M} \left( 1 - \prod_{i \in \partial a} \delta_{x_i, J_{ia}} \right) \right]. \quad (3)$$

We define a large deviation function $\mathcal{L}(s)$ as the logarithm of this number divided by the size of the system $N$. Again to compute $\mathcal{L}(s)$ it is advantageous to introduce its Legendre transform

$$Z(x) = e^{N\Phi(x)} = \sum_\mathcal{J} e^{x N s(\mathcal{J})} = e^{N[\mathcal{L}(s)+xs]}, \quad \frac{\partial \mathcal{L}(s)}{\partial s} = -x, \quad (4)$$

where the saddle-point approximation for $N \to \infty$ has been used.

We stress here that $\mathcal{L}(s)$ is the large deviation function with respect to the negation-configurations; it is hence defined for a given geometry of the satisfiability formula. In what follows we assume, as is usual, that $\mathcal{L}(s)$ is self-averaging with respect to the
Adversarial satisfiability problem

formula geometry, i.e. \( L(s) \) is almost surely the same function for two randomly chosen formulas. This assumption is a generalization of the self-averaging property of the free energy in the canonical \( K \)-SAT problem. Note also that when writing this expression we implicitly assume that the number of negation-configurations that give a certain entropy is exponential in \( N \). If it is smaller that exponential in \( N \) computation of \( \Phi(x) \) will lead to \( L(s) = -\infty \). We will come back to this point in section 6.

Note two special cases: for \( x = 0 \) the partition function (4) is simply equal to the total number of negation-configurations \( \Phi(0) = K \alpha \log 2 \); for \( x = 1 \), the partition function above is related to the annealed partition function, \( \Phi(1) = \log 2 + \alpha \log(2^K - 1) \).

The major difficulty in calculating \( \Phi(x) \), for a general value of \( x \), is that the entropy \( s(J) \) is not defined as a sum of local terms. On the other hand the geometry of the underlying factor graph is still tree-like in the random AdSAT, hence for any configuration of negations \( J \) we can apply the cavity method (with replica symmetry breaking if needed) to compute the entropy \( s(J) \). In the cavity method, as is reminiscent of the Bethe approximation, the entropy (or more generally Bethe free energy) can be written as a sum of local terms. This fact enables us to calculate \( \Phi(x) \).

The statistical physics treatment of the random \( K \)-SAT problem among others led to a discovery that the replica symmetry breaking approach is needed \([22,8,9]\) in order to correctly compute the entropy close to the satisfiability threshold \( \alpha_c \). In other words, in that region the space of solutions splits into well ergodically separated clusters. We define the complexity function \( \Sigma \) as the logarithm of the total number of clusters per variable. The value of the complexity can then be computed with the survey propagation equations \([8]\). At the satisfiability threshold the complexity goes to zero, whereas the entropy density of solutions is a positive number even at the threshold. With this in mind it will be useful to define also

\[
e^{N\Phi_{SP}(x)} = \sum_J e^{xN\Sigma(J)} = e^{N[L_{SP}(\Sigma) + x\Sigma]}, \quad \frac{\partial L_{SP}(\Sigma)}{\partial \Sigma} = -x,
\]

where \( L_{SP}(\Sigma) \) is the entropy density of negation-configurations that give a certain complexity function \( \Sigma \), and the saddle-point approximation for \( N \to \infty \) has been used.

3. Reminder of equations for belief and survey propagation

With the notation introduced in section 2 we write the belief propagation equations and the Bethe entropy as derived, e.g., in \([23,24]\). These equations are asymptotically exact on locally tree-like graphs as long as all the correlation length scales are finite. If they are not then splitting the phase space into clusters such that within each cluster the correlations decay again might be possible. SP then estimates the total number of such clusters \([8]\), and it does so asymptotically exactly at least close enough to the satisfiability threshold \([25,9]\).

Denoting by \( \{m_{ia}, \hat{m}_{ai}\} \) the BP (SP) messages, we write the BP (SP) fixed-point equations as

\[
m_{ia} = g_{ia}(\{\hat{m}_{bi}\}_{b \in \partial a \setminus a}, \{J_{bi}\}_{b \in \partial i}), \quad (6)
\]

\[
\hat{m}_{ai} = \hat{g}_{ai}(\{m_{ja}\}_{j \in \partial a \setminus i}, \{J_{ja}\}_{j \in \partial a}), \quad (7)
\]
In the BP case, the messages read \([24]\) \(m_{ia} = \{\nu_{ia}^0, \nu_{ia}^1\}, \hat{m}_{ai} = \{\hat{\nu}_{ai}^0, \hat{\nu}_{ai}^1\}\), and

\[
\begin{align*}
g_{ia}^r(\{\nu_{ba}\}_{b \in \partial i \setminus a}) &= \frac{\prod_{b \in \partial i \setminus a} \nu_{ba}^r \nu_{ba}^1}{\prod_{b \in \partial i \setminus a} \nu_{ba}^1 + \prod_{b \in \partial i \setminus a} \nu_{ba}^r}, \\
\hat{g}_{ai}^r(\{\nu_{ja}\}_{j \in \partial a \setminus i}, \{J_{ja}\}_{j \in \partial a}) &= \frac{1 - \delta_{r, J_{ai}} \prod_{j \in \partial a \setminus i} \nu_{ja}^{J_{ja}}}{2 - \prod_{j \in \partial a \setminus i} \nu_{ja}^{J_{ja}}},
\end{align*}
\]

where \(r = 0, 1\). In the SP case, \(m_{ia} = \{Q_{ia}^S, Q_{ia}^U, Q_{ia}^*\}, \hat{m}_{ai} = \hat{Q}_{ai}\), and

\[
\begin{align*}
g_{ia}^s(\{J_{ba}\}_{b \in \partial i}, \{\hat{Q}_{ba}\}_{b \in \partial i \setminus a}) &= C \prod_{b \in \partial i \setminus a} (1 - \hat{Q}_{ba}), \\
g_{ia}^S(\{J_{ba}\}_{b \in \partial i}, \{\hat{Q}_{ba}\}_{b \in \partial i \setminus a}) &= C \prod_{b \in \mathcal{U}_{ia}} (1 - \hat{Q}_{ba}) \left[ 1 - \prod_{b \in \mathcal{S}_{ia}} (1 - \hat{Q}_{ba}) \right], \\
g_{ia}^U(\{J_{ba}\}_{b \in \partial i}, \{\hat{Q}_{ba}\}_{b \in \partial i \setminus a}) &= C \prod_{b \in \mathcal{S}_{ia}} (1 - \hat{Q}_{ba}) \left[ 1 - \prod_{b \in \mathcal{U}_{ia}} (1 - \hat{Q}_{ba}) \right], \\
\hat{g}_{ai}(\{Q_{ja}\}_{j \in \partial a \setminus i}) &= \prod_{j \in \partial a \setminus i} Q_{ja}^U,
\end{align*}
\]

where \(C\) is a normalization constant enforcing the relation \(g_{ia}^s + g_{ia}^S + g_{ia}^U = 1\), and \(\mathcal{S}_{ia}, \mathcal{U}_{ia}\) are defined as

\[
\begin{align*}
\text{if } J_{ia} = 0 & \quad \mathcal{S}_{ia} = \partial_{0i} \setminus a, \quad \mathcal{U}_{ia} = \partial_{1i}, \\
\text{if } J_{ia} = 1 & \quad \mathcal{S}_{ia} = \partial_{1i} \setminus a, \quad \mathcal{U}_{ia} = \partial_{0i},
\end{align*}
\]

(10)

where \(\partial_{0/1i} = \{a \in \partial i\} \text{ such that } J_{ia} = 0/1\).

If \(\{m_{ia}, \hat{m}_{ai}\}\) is a fixed point of equations (6)–(7), the Bethe entropy for BP and the complexity for SP are both written in a general form

\[
s(\{J_{ia}, m_{ia}, \hat{m}_{ai}\}) = \sum_{a=1}^{M} s_a(\{m_{ia}, J_{ia}\}_{i \in \partial a}) + \sum_{i=1}^{N} s_i(\{\hat{m}_{ai}, J_{ia}\}_{a \in \partial i}) - \sum_{(ia)} s_{ia}(m_{ia}, \hat{m}_{ai}),
\]

(11)

where for BP

\[
\begin{align*}
s_a(\{\nu_{ia}, J_{ia}\}_{i \in \partial a}) &= \log \left( 1 - \prod_{i \in \partial a} \nu_{ia}^{J_{ia}} \right), \\
s_i(\{\hat{\nu}_{ai}, J_{ia}\}_{a \in \partial i}) &= \log \left( \prod_{b \in \partial i} \hat{\nu}_{bi}^0 + \prod_{b \in \partial i} \hat{\nu}_{bi}^1 \right), \\
s_{ia}(\nu_{ia}, \hat{\nu}_{ai}) &= \log \left( \nu_{ia}^0 \hat{\nu}_{ai}^0 + \nu_{ia}^1 \hat{\nu}_{ai}^1 \right),
\end{align*}
\]

(12)
while for SP
\[
S_a(\{Q_{ia}, J_{ia}\}_{ia}) = \log \left( 1 - \prod_{j \in \partial a} Q_{ja}^U \right).
\]
\[
S_i(\{\hat{Q}_{ai}, J_{ai}\}_{ai}) = \log \left[ \prod_{b \in \partial ai} (1 - \hat{Q}_{bi}) + \prod_{b \in \partial i} (1 - \hat{Q}_{bi}) - \prod_{b \in \partial i} (1 - \hat{Q}_{bi}) \right],
\]
\[
S_{ai}(Q_{ia}, \hat{Q}_{ai}) = \log \left( 1 - Q_{ia}^U \hat{Q}_{ai} \right).
\]

Readers unfamiliar with the interpretation and derivation of the belief and the survey propagation (8)–(10) are referred to [23, 26, 24]. However, for understanding of our method in what follows the general form (6)–(7) and (11) is sufficient.

4. Computation of the large deviations function

The most important formula of section 3 is (11): in certain regimes it gives the asymptotically exact entropy or complexity in a form factorized in local terms. The remaining complication is that now everything depends on the fixed point of the BP (SP) equations. We can, however, write
\[
Z(x) = \sum_\mathcal{J} \int \prod_{ia} dm_{ia} d\hat{m}_{ia} e^{Nxs(\{J_{ia}, m_{ia}, \hat{m}_{ai}\})} \prod_{(ia)} \delta(m_{ia} - g_{ia}(\{\hat{m}_{ia}\}_{b \in \partial \hat{a}}, \{J_{ia}\}_{b \in \partial a})) \\
\times \prod_{(ia)} \delta(\hat{m}_{ai} - \hat{g}_{ai}(\{m_{ja}\}_{j \in \partial a \setminus i}, \{J_{ja}\}_{j \in \partial a})).
\]

If we now introduce auxiliary variables \(\omega_{ia} \equiv \{J_{ia}, m_{ia}, \hat{m}_{ai}\}\) the partition function defined by (4) can be re-written in the common local form
\[
Z(x) = \sum_{\{\omega_{ia}\}} \left\{ \prod_{a=1}^M \Psi_a(\{\omega_{ia}\}_{i \in \partial a}) \right\} \left[ \prod_{i=1}^N \Psi_i(\{\omega_{ia}\}_{a \in \partial i}) \right] \left[ \prod_{(ia)} \Psi_{ai}(\omega_{ia}) \right],
\]
where
\[
\Psi_a(\{\omega_{ia}\}_{i \in \partial a}) \equiv e^{xS_a(\{m_{ia}, J_{ia}\}_{i \in \partial a})} \prod_{i \in \partial a} \delta(\hat{m}_{ai} - \hat{g}_{ai}(\{m_{ja}\}_{j \in \partial a \setminus i}, \{J_{ja}\}_{j \in \partial a})),
\]
\[
\Psi_i(\{\omega_{ia}\}_{a \in \partial i}) \equiv e^{xS_i(\{\hat{m}_{ai}, J_{ai}\}_{a \in \partial i})} \prod_{a \in \partial i} \delta(m_{ia} - g_{ia}(\{\hat{m}_{ia}\}_{b \in \partial \hat{a} \setminus i}, \{J_{ia}\}_{b \in \partial a})),
\]
\[
\Psi_{ai}(\omega_{ia}) \equiv e^{-xS_{ai}(m_{ia}, \hat{m}_{ai})}.
\]

In equation (15) and in the following the sum over \(\omega_{ia}\) stands for the sum over \(J_{ia}\) and the integral over \(m_{ia}, \hat{m}_{ai}\). The probability measure in equation (15) is local and can hence be represented with an auxiliary factor graph that can be viewed as decorating the original K-SAT factor graph. Figure 1 depicts this construction.

The partition function \(Z(x)\) can now be computed by implementing the general BP formalism to the auxiliary graph, just as is done in the derivation of the 1RSB equations in [24] (note indeed the close formal resemblance of our approach and the 1RSB equations).
Adversarial satisfiability problem

Figure 1. Left: the graph of the original $K$-SAT instance. Empty circles and empty squares represent variable and function nodes respectively. An edge connecting a variable node $i$ to a function node $a$ means that the function node $\psi_a$ depends on $x_i$. Right: the auxiliary factor graph describing equation (15), built upon the graph of the $K$-SAT instance in the left panel. Empty squares with black-filled squares inside represent the $\Psi_a$ function nodes, empty circles with black-filled squares inside represent $\Psi_i$ function nodes of the auxiliary graphs. Finally, black-filled squares represent $\Psi_i a$ function nodes, while black-filled circles represent $\omega_i a$ variable nodes of the auxiliary graph. An edge connecting $\omega_i a$ to a function node of the auxiliary graph means that such a function node depends on $\omega_i a$.

We call $S_{ia}(\omega_{ia})$ the message going from the variable node $ia$ to the function node $a$, and $\hat{S}_{ai}(\omega_{ia})$ the message going from the variable node $ia$ to the function node $i$. BP equations on the auxiliary factor graph on the variables $\omega_{ia}$ then lead to fixed-point equations for these messages:

$$\hat{S}_{ai}(\omega_{ia}) \simeq \sum_{\{\omega_{ja}\}_{j \in \partial a \setminus i}} \left[ \hat{z}_{ai}\left(\{m_{ja}, J_{ja}\}_{j \in \partial a}, \hat{m}_{ai}\right) \right]^x \prod_{j \in \partial a} \delta(\hat{m}_{aj} - \hat{g}_{aj}\left(\{m_{ka}\}_{k \in \partial a \setminus j}, \{J_{ka}\}_{k \in \partial a}\right)) \times \prod_{j \in \partial a \setminus i} S_{ja}(\omega_{ja}), \quad (17)$$

$$S_{ia}(\omega_{ia}) \simeq \sum_{\{\omega_{ib}\}_{b \in \partial i \setminus a}} \left[ z_{ia}\left(\hat{m}_{bi}, J_{ib}\right)_{b \in \partial i}, m_{ia}\right]^x \prod_{b \in \partial i} \delta(m_{ib} - g_{ib}\left(\hat{m}_{ci}\right)_{c \in \partial i \setminus b}, \{J_{ci}\}_{c \in \partial i}) \times \prod_{b \in \partial i \setminus a} \hat{S}_{bi}(\omega_{ib}), \quad (18)$$

where

$$z_{ia}\left(\hat{m}_{bi}, J_{ib}\right)_{b \in \partial i}, m_{ia}\right) \equiv e^{S_i(\hat{m}_{bi}, J_{ib})_{b \in \partial i} - S_{ia}(m_{ia}, \hat{m}_{ai})},$$

$$\hat{z}_{ai}\left(\hat{m}_{ja}, J_{ja}\right)_{j \in \partial a}, \hat{m}_{ai}\right) \equiv e^{S_a(\hat{m}_{ja}, J_{ja})_{j \in \partial a} - S_{ai}(m_{ia}, \hat{m}_{ai})}.$$

Equations (17)–(18) can be further simplified. It is easy to check that when the fixed-point equations (6)–(7) hold, the term $z_{ia}$ (respectively $\hat{z}_{ai}$) does not depend on the ‘backward’ messages $m_{ia}$ (respectively $\hat{m}_{ai}$). Using this result, we can see that (17)–(18) are compatible with a choice of the messages $S_{ia}$, $\hat{S}_{ai}$ depending only on $m_{ia}, J_{ia}$ and only on $\hat{m}_{ai}, J_{ia}$ respectively. Indeed, if we assume that $S_{ia}(\omega_{ia}) = S_{ia}(m_{ia}, J_{ia})$, and
\[ \dot{S}_{ia}(\omega_{ia}) = \dot{S}_{ia}(\hat{m}_{ia}, J_{ia}) \text{, then } (17) \text{–}(18) \text{ become} \]

\[ S_{ia}(m_{ia}, J_{ia}) \simeq \sum_{\{J_{ia}\} \in \partial a} \int \prod_{b \in \partial a \setminus a} \delta \hat{m}_{ib} \hat{S}_{ib}(\hat{m}_{ib}, J_{ib}) [z_{ia}(\{\hat{m}_{ib}, J_{ib}\}_{b \in \partial a \setminus a})]^x \times \delta (m_{ia} - g_{ia}(\{\hat{m}_{ib}, J_{ib}\}_{b \in \partial a \setminus a}, \{J_{ia}\}_{b \in \partial i})) \]

\[ \dot{S}_{ia}(\hat{m}_{ia}, J_{ia}) \simeq \sum_{\{J_{ia}\} \in \partial a \setminus i} \int \prod_{j \in \partial a \setminus i} dm_{ja} S_{ja}(m_{ja}, J_{ja}) [\hat{z}_{ia}(\{m_{ja}, J_{ja}\}_{j \in \partial a \setminus i})]^x \times \delta (\hat{m}_{ia} - \hat{g}_{ia}(\{m_{ja}, J_{ja}\}_{j \in \partial a \setminus i}, \{J_{ia}\}_{j \in \partial a})) \]  

(19)

(20)

The free energy \( \Phi(x) \) can then be computed using the general expression for the Bethe free entropy [24]

\[ N \Phi(x) = \sum_{a=1}^{M} F_a + \sum_{i=1}^{N} F_i - \sum_{(ia)} F_{ia}, \]  

(21)

where

\[ F_a = \log \left[ \sum_{\{J_{ia}\} \in \partial a} \int \prod_{i \in \partial a} dm_{ia} S_{ia}(m_{ia}, J_{ia}) e^{z_{ia}(\{m_{ia}, J_{ia}\}_{i \in \partial a})} \right], \]  

(22)

\[ F_i = \log \left[ \sum_{\{J_{ia}\} \in \partial i} \int \prod_{a \in \partial i} dm_{ia} \hat{S}_{ia}(\hat{m}_{ia}, J_{ia}) e^{\hat{z}_{ia}(\{\hat{m}_{ia}, J_{ia}\}_{a \in \partial i})} \right], \]  

(23)

\[ F_{ia} = \log \left[ \sum_{J_{ia}} \int dm_{ia} dm_{ia} S_{ia}(m_{ia}, J_{ia}) \hat{S}_{ia}(\hat{m}_{ia}, J_{ia}) e^{z_{ia}(m_{ia}, \hat{m}_{ia})} \right]. \]  

(24)

Equations (19)–(20) clearly show the formal analogy with the 1RSB cavity equations [24]. The difference between equations (19)–(20) and the latter is that in the AdSAT case the negations are considered as physical degrees of freedom of the partition function \( Z(x) \), and the resulting cavity equations (19)–(20) consist in a weighted average of the quantities \([z_{ia}(\{\hat{m}_{ib}, J_{ib}\}_{b \in \partial a \setminus a})]^x, [\hat{z}_{ia}(\{m_{ja}, J_{ja}\}_{j \in \partial a \setminus i})]^x\) over \( \{J_{ia}, m_{ia}, \hat{m}_{ia}\} \). On the contrary, in the 1RSB case the only degrees of freedom are the BP messages \( \{m_{ia}, \hat{m}_{ia}\} \) in such a way that the resulting 1RSB cavity equations consist in an average over \( \{m_{ia}, \hat{m}_{ia}\} \) at fixed \( J_{ia} \)s.

The biggest advantage of the formal resemblance of equations (19)–(20) to the 1RSB cavity equations is that in order to solve (19)–(20) numerically we can use the very same technique and all the related knowledge as in the case of 1RSB. We indeed use the population dynamics [22], where the distributions \( S_{ia}(m_{ia}, J_{ia}), \hat{S}_{ia}(\hat{m}_{ia}, J_{ia}) \) are represented as populations of \( P \) messages. When the size of the population \( P \) is large, we expect the populations to reproduce well the distributions \( S_{ia}(m_{ia}, J_{ia}), \hat{S}_{ia}(\hat{m}_{ia}, J_{ia}) \).

The cavity equations (19)–(20) can be written in terms of these populations. Starting from a given initial configuration, the iteration of the cavity equations yields the fixed-point populations satisfying (19)–(20). Once this fixed point is achieved the free energy \( \Phi(x) \) can be computed numerically by means of (21)–(24). This is repeated for different values
of $x$ and finally the Legendre transform $\mathcal{L}(s)$ is evaluated. Everything is done in the very same way as the 1RSB equations are usually solved, for more details see, e.g., [22, 27, 28]. The only difference is in the treatment of the negation-variables. In our case one writes the distributions $S_{ia}(m_{ia}, J_{ia})$, $\hat{S}_{ai}(\hat{m}_{ai}, J_{ia})$ as

$$S_{ia}(m_{ia}, J_{ia}) = \frac{1}{2} S_{ia}(m_{ia}, J_{ia}), \quad \hat{S}_{ai}(\hat{m}_{ai}, J_{ia}) = \frac{1}{2} \hat{S}_{ai}(\hat{m}_{ai}, J_{ia}),$$

because $S_{ia}(J_{ia}) = \int dm_{ia} S_{ia}(m_{ia}, J_{ia}) = 1/2 = \hat{S}_{ai}(J_{ia})$, as can be seen by explicitly integrating equations (19)–(20). We then introduce a pair of populations $\{S_{ia}^1[s], \hat{S}_{ai}^1[s]\}_{s=1,\ldots,P}$ and $\{S_{ia}^0[s], \hat{S}_{ai}^0[s]\}_{s=1,\ldots,P}$ representing the probability distributions $S_{ia}(m_{ia}|1)$, $\hat{S}_{ai}(\hat{m}_{ai}|1)$ and $S_{ia}(m_{ia}|0)$, $\hat{S}_{ai}(\hat{m}_{ai}|0)$ respectively. The population dynamics is then implemented in terms of such populations, and the resulting fixed point investigated numerically.

5. Cavity method results for random AdSAT

In this section we present the solution of the cavity equations (19)–(20), and its implications for the random AdSAT problem.

5.1. Large deviations of the entropy and complexity on regular instances

Before addressing the random AdSAT as defined in section 1 we will study it on random regular instances. On $L$-regular instances every variable belongs to exactly $L$ clauses. A random $L$-regular instance is chosen uniformly at random from all possible ones with a given number of variables $N$ and number of clauses $M$, provided that $KM = LN$. Note that, as far as we know, this ensemble of random SAT instances was not treated in the literature previously.

In the 3-SAT problem discussed in this paper the Bethe entropy is asymptotically exact only as long as the BP equations converge to a fixed point [9]; the non-convergence is equivalent to the spin glass instability, or a continuous transition to a replica symmetry breaking phase. On random regular graphs with random values of negations, BP stops converging at $L = 12$ (this is the reason why these and larger values are omitted from table 1), meaning that for $L \geq 12$ there is a need for SP (or another form of replica symmetry breaking solution). For $L \geq 15$ BP iterations for random regular 3-SAT lead to contradictions (zero normalizations) meaning that in this region the large random instances are almost surely unsatisfiable. Survey propagation on random regular 3-SAT has a trivial fixed point for $L \leq 12$, and a non-trivial fixed point for $L = 13$ with the value of complexity $\Sigma(L = 13) = 0.008$. SP does not converge for $L \geq 14$; if we ignore the non-convergence, and compute the complexity from the current values of messages, we get on average $\Sigma(L = 14) = -0.03$. This means that $L = 13$ is the largest satisfiable case.

The great advantage of random regular instances is that topologically the local neighborhood looks the same for every variable $i$. Moreover, we recall that in regimes where the BP equations are asymptotically exact the properties of variable $i$ depend only on the structure of the local neighborhood of $i$. Hence on regular graphs all the quantities in equations (19)–(20) are independent of the indices $i, j$, and $a, b$. This so-called factorization property simplifies crucially the numerical solution of
Table 1. The Bethe free entropy on regular instances with random negation-configurations ($s_{\text{ran}}$), with balanced negation-configurations ($s_B$), and non-frustrated with $J_{ia} = 0 \forall (ia)$ ($s_U$). The entropy for the non-frustrated case and for the balanced case for even degree $L \leq 10$ can be computed analytically since the BP fixed point is factorized in these cases. In the other cases we iterate the BP equations on large random graphs and compute the entropy from the corresponding BP fixed point. The star signals that BP did not converge and the value of entropy was obtained by averaging over an interval of time. The $\times$ means that BP converged to contradictions for these densities of constraints.

| $L$ | $s_{\text{ran}}$ | $s_B$ | $s_U$ |
|-----|------------------|-------|-------|
| 2   | 0.6039           | 0.5710| 0.6196|
| 3   | 0.5592           | 0.5324| 0.5975|
| 4   | 0.5134           | 0.4488| 0.5796|
| 5   | 0.4686           | 0.4120| 0.5644|
| 6   | 0.4220           | 0.3266| 0.5513|
| 7   | 0.3750           | 0.2902| 0.5397|
| 8   | 0.3302           | 0.2044| 0.5293|
| 9   | 0.2816           | 0.1677| 0.5199|
| 10  | 0.2319           | 0.082 *| 0.5114|
| 11  | 0.1813           | 0.042* | 0.5035|
| 12  | 0.128 * $\times$|       | 0.4962|
| 13  | 0.07 * $\times$  |       | 0.4894|
| 14  | $\times$         |       | 0.4831|

equations (19)–(20); the $2KM$ distributions $\{S_{ia}(\nu_{ia},J_{ia}),\hat{S}_{ai}(\hat{\nu}_{ai},J_{ai})\}$ reduce to only two ($J = 0$ and 1) distributions $S(\nu,J),\hat{S}(\hat{\nu},J)$. Moreover, the thermodynamic limit is taken directly without increasing the computational effort. (To avoid confusion, we recall here that for the canonical random $K$-SAT problem where negation-variables are chosen uniformly at random and fixed, the BP solution is not factorized. In the adversarial version one sums over the negation-variables, hence the factorization.)

In figure 2 we show the large deviation function $\mathcal{L}(s)$ of the Bethe entropy $s$ obtained by the population dynamics over BP messages on regular graphs with $K = 3$ and variable degree $L = 4$, i.e. by solving equations (19), (20) and (21)–(24). First of all, in the ‘infinite temperature’ case, i.e. when the Legendre parameter $x = 0$, that is at the maximum of $\mathcal{L}(s)$, we recover the logarithm of the total number of negation-configurations $\mathcal{L}(x = 0) = L \log 2$. The corresponding value of entropy $s(x = 0) = s_{\text{ran}}$ is the Bethe entropy for a random choice of negations (values summarized in table 1).

The inset of the figure shows that as the Legendre parameter $x \to \pm \infty$ both $\mathcal{L}$ and $s$ converge to well defined ending points (the same data in a logarithmic plot show that the convergence is exponential). Let us denote the lowest entropy ending point (left, $x \to -\infty$) ($s_L,\mathcal{L}_L$), and the highest entropy ending point (right, $x \to \infty$) ($s_R,\mathcal{L}_R$). We observe systematically that the value of $s_R$ is equal to $s_U$, where $s_U$ is the entropy of the uniform negation-configuration which is obtained by computing a fixed point of equations (6)–(7) such that $\hat{\nu}_{ai} = \hat{\nu} \forall (ai)$, $\nu_{ia} = \nu \forall (ia)$ and $J_{ia} = 0 \forall (ia)$, and plugging it into equation (11). The values of $s_U$ as a function of $L$ are summarized in table 1. An edge-independent fixed point of the BP equations is called factorized. We realize that $s_U$
Figure 2. Left: the BP large deviation function $\mathcal{L}(s)$ versus the Bethe entropy $s$ computed by population dynamics on regular graphs with $K = 3, L = 4$, population size $P = 10^4$. The left ending point $(f_L, L_L)$ corresponds to balanced configurations of negations, whereas the right ending point $(f_R, L_R)$ to the polarized configurations of negations (more details in the text). Right: the SP large deviation function $\mathcal{L}_{SP}(\Sigma)$ versus the complexity $\Sigma$ computed by population dynamics on regular graphs with $K = 3, L = 13, P = 10^4$ (bottom), and $K = 3, L = 12, P = 75 \times 10^3$ (top).

also corresponds to the value of the Bethe entropy when every variable is either always negated or never negated; we call such negation-configurations polarized. There are $2^{L_N}$ polarized negation-configurations, and indeed the logarithm of the number of such choices corresponds to the value of $L_R = \log 2$. Intuitively such configurations of negation are frustrating the formula in the least possible way, and figure 2 shows that such intuition is asymptotically exact in this case.

Similarly, for the lowest entropy ending point, for even values of the degree $L$, we realize that $L_L = \log \left( \frac{L}{L/2} \right)$ and $s_L$ corresponds to a value $s_B$ that is obtained from a factorized BP fixed point when each variable is $L/2$ times negated and $L/2$ times non-negated; values are summarized in table 1. Such balanced configurations of negations locally frustrate the variables in a maximal way (half of the clauses want the variable to be 1, the other half 0). And the computation presented in figure 2 suggests that asymptotically there are no correlated negation-configurations that would frustrate the formula even more and decrease the value of the entropy further.

We investigate in more detail the result following from figure 2, i.e. that the most frustrated configurations of negations on the regular graphs with even degree are the balanced negations; we denote the balanced negation-configurations $\{J_{ia}\}_B = J_B$. There are $\left( \frac{L}{L/2} \right)^N$ such negation-configurations. Does our result mean that all of them lead to the same number of solutions $N(J_B)$? We will see in section 6 that this is not true for finite $N$. The correct conclusion from the result presented in figure 2 is that $\lim_{N \to \infty} [\log N(J_B)]/N = s_B = s_L$ independently of the realization of $J_B$. This can also be seen directly from the solution of the BP equations on the formulas with balanced negations. Indeed, for even $L$ the fixed point of the BP equations is factorized and

doi:10.1088/1742-5468/2011/03/P03023
Adversarial satisfiability problem independent of the realization of negations and also of the size of the graph. We tried numerically formulas of various sizes and many possible realizations of balanced negations, and for even degree $L < 10$ BP always converges to the factorized fixed point (at $L = 10$ BP stops converging, as we will discuss later in the paper), giving always the same Bethe entropy density $s_B$. Further discussion about the true entropy fluctuations compared to the constant Bethe entropy in this case will be presented in section 6.

For regular graphs with odd degree $L$ we cannot achieve ideal balancing of every variable. Instead, we call a configuration of negations balanced if for every variable there is either $(L - 1)/2$ or $(L + 1)/2$ negations. The total number of such configurations is then $2^N(L - 1)/2 N$. The BP fixed point on the balanced instances for odd $L$ is not factorized anymore. We can, however, solve the cavity equations (19)–(20) restricted to only balanced values of negations and we obtain that within the error-bars of the numerical resolution of the equations (that are less than 1%) all the balanced configurations give the same value of Bethe entropy also in the odd $L$ case.

Our results for large deviations of the Bethe entropy lead to a conclusion that for the regular instances and in the limit $N \to \infty$ the most frustrated formulas are all those with balanced configurations of negations. Let us hence conclude this section by summarizing the properties of regular SAT instances with balanced negations. BP on balanced instances converges for $L \leq 9$, and leads to contradictions for $L \geq 13$. Survey propagation on balanced regular instances has a trivial fixed point for $L \leq 9$, for $L = 10$ a fixed point with complexity $\Sigma_B(L = 10) = 0.018$, and for $L \geq 11$ the complexity is negative (e.g. $\Sigma_B(L = 11) = -0.001$, $\Sigma_B(L = 12) = -0.075$).

For completeness, we also computed the large deviations of the complexity function. That is, we solved equations (19)–(20) using SP as the basic message passing scheme. Figure 2 (right) shows some of the results for $L = 12$ and 13; we indeed see that there are configurations of negations that lead to negative complexity. Unfortunately, it is hard to extract any information from these curves for very negative values of $x$, because of the noise introduced by the finite population size effects. This also poses a problem for $L = 10$ and 11, where we know that a non-trivial fixed point of SP exists for the balanced configurations of negations. In the population dynamics we should hence see a non-trivial solution for very negative values of $x$. Instead we were only able to obtain very noisy and inconclusive data from the population dynamics with population sizes up to $7.5 \times 10^4$. For $L = 10$ the SP equations have only one factorized fixed point for all the balanced configurations of negations; this again strongly suggests that instances with balanced negations are the most frustrated ones, and hence that for $L = 10$ the adversary cannot make large formulas unsatisfiable. For lower values of $L \leq 9$ the population dynamics has always only a trivial fixed point given by $S_{ia}(Q_{ia}, J_{ia}) = \delta(Q^S_{ia})\delta(Q^U_{ia})\delta(Q^*_{ia} - 1)/2$, $\hat{S}_{ai}(\hat{Q}_{ai}, J_{ia}) = \delta(\hat{Q}_{ai})/2$ yielding $\Phi(x) = K\alpha \log 2$. SP is hence not very useful in this case to obtain new information about the random AdSAT problem.

In summary, for $L \geq 11$ the adversary will succeed in making a large formula unsatisfiable by simply balancing the negations (for $L \geq 14$ a random choice of negations would do). On the other hand, following our previous conclusion that the balanced formulas are the most frustrated ones, for $L \leq 10$ the adversary will not be able to make large random regular SAT instances unsatisfiable by adjusting the values of negation-variables.
5.2. Results for random AdSAT, i.e. instances with Poisson degree distribution

In the most commonly considered ensemble of the random $K$-satisfiability problem, $K$ variables appearing in each clause are chosen independently at random (avoiding repetitions). For large system sizes this procedure generates Poissonian degree distribution with mean $K\alpha$. In this case every node has a different local neighborhood and hence the fixed point of equations (19)–(20) is not factorized, and the distribution $S_{ia}(m_{ia}, J_{ai})$, $\hat{S}_{ia}(\hat{m}_{ia}, \hat{J}_{ai})$ are different on every edge. We hence solve equations (19)–(20) by generating an instance of the problem (graph) of size $N$, associating one population of size $P$ with every directed edge, and iterate following equations (19)–(20). This is more computationally involved, and we are able to treat only modestly large $N$ and $P$, typically several hundreds. The resulting large deviation function $L(s)$ is depicted in figure 3 for several values of constraint density $\alpha$.

For low values of the constraint density, e.g. $\alpha = 1$ in figure 3, the location of the right (large entropy, least frustrated) ending point $(s_R, L_R)$ corresponds, as in the case of random regular instances, to the value of Bethe entropy that is obtained if no negations are present in the instance ($J_{ia} = 0$ for all $ia$), and $L_R = (1 - e^{-K\alpha}) \log 2$ (corresponding to the number of negation-configurations where no variable is locally frustrated). For larger values of the constraint density, e.g. $\alpha = 2$ in figure 3, the results from population sizes as large as we were able to achieve are very noisy for large values of $x \approx 100$. We observed that the data are getting smoother as the population size is growing, however not enough to be able to confirm from these data that $(s_R, L_R)$ is the right ending point.

The part of the curve corresponding to a very large negative parameter $x$ does not converge to an ending point. Instead at some $x_0$ the large deviation function $L(s)$ ceases to be concave, and an unphysical branch appears for $x < x_0$. This unphysical branch is not present on random regular instances with even degree; when the degree is odd the data for large negative $x$ are inconclusive in the sense that we might see a unphysical branch or only a numerical noise. We define the left ending point as the extreme of the physical branch $s_L = s(x_0)$, and $L_L = L(x_0)$. We observe systematically that in the region of interest (say for $\alpha \geq 1$) the values $s_L$ and $L_L$ are very close to the values corresponding to balanced instances $(s_B, L_B)$. In balanced instances each variable is negated as many times as non-negated (for variables of odd degree the absolute value of the difference between the number of negations and non-negations is one). In the thermodynamic limit the number of such balanced negation-configurations is

$$L_B = \sum_{i=0}^{\infty} \log \left( \frac{2i}{i} \right) e^{-k\alpha} \frac{(k\alpha)^{2i}}{(2i)!} + \sum_{i=0}^{\infty} \log \left[ 2 \left( \frac{2i+1}{i} \right) \right] e^{-k\alpha} \frac{(k\alpha)^{2i+1}}{(2i+1)!}. \quad (25)$$

We made a number of attempts to obtain a value of entropy considerably smaller than the balanced entropy, $s < s_B$. First, we removed the leaves from the formula and balanced only the residual formula. This indeed leads to a lower value of the entropy, but for $\alpha > 1$ the difference was less than 1%. We tested the population dynamics limited to the balanced negation-configurations, i.e. we solved equations (19)–(24) where the sum over the negation-variables in equations (20) and (23) was limited only to the balanced negation-configurations. The large deviation function $L(s)$ obtained in this way did not differ more than by 1% from the value $(s_B, L_B)$, see figure 3 (left). We also investigated the results of population dynamics over the SP equations and we were not able to find cases...
Figure 3. Left: the BP large deviation function $\mathcal{L}(s)$ versus the Bethe entropy $s$ computed by population dynamics on Poissonian graphs with random negation-configurations, for $K = 3$, various values of the constraint density $\alpha$, and both positive and negative $x$. For $\alpha = 1–3$ the part of the curve with $x < 0$ has been computed with $N = P = 300$. For $\alpha = 1$ the part of the curve with $x > 0$ reaches the right ending point, and has been computed with $N = P = 500$. For $\alpha = 2, 3$ the part of the curve with positive $x$ has been computed with $N = P = 300$, and it does not reach the right ending point; even larger $N$ and $P$ are needed to remove the noise from the data for very large positive $x$. For $x < 0$, an unphysical branch (concave part of the curve) starts at $x_0 \approx -42, x_0 \approx -44, x_0 \approx -40$ for $\alpha = 1, 2, 3$ respectively. The points indicate the values for balanced configurations of negations ($s_B, \mathcal{L}_B$), and for configurations of negations where all the variables are non-negated ($s_R, \mathcal{L}_R$). Right: zoom of the large deviation function $\mathcal{L}(s)$ versus $s$ for $\alpha = 1$ close to the low entropy ending point for $N = P = 500$. The data become noisy close to the low entropy ending point; larger graph and population sizes lead to an improvement. We plotted the data down to the lowest value of entropy $s$; hence the unphysical branch is not plotted. The point indicates the value ($s_B, \mathcal{L}_B$) for balanced configurations of negations. The blue data points show the large deviation function restricted to balanced configurations of negations for $N = 10^2, P = 10^3$. Notice the narrow range of entropies $s$ plotted in the latter, and how little the lowest entropy we achieved differs from the balanced value.

where the complexity would decrease by more than 1% below the complexity value on the balanced instances. We also tried simulated annealing on the negation-variables using the Bethe entropy as the cost function, with the same result. All this makes us conclude that with at least 1% of precision the satisfiability threshold for random adversarial SAT equals the satisfiability threshold of the balanced random ensemble.

Let us hence summarize the results about BP and SP for the random satisfiability problem with balanced configurations of negations. For $K = 3$ the BP ceases to converge for $\alpha \geq 2.96$. SP starts to converge to a non-trivial fixed point for $\alpha > 3.20$, and the complexity decreases to zero at

$$\alpha_B = 3.399 \pm 0.001;$$  \hspace{1cm} (26)

this is hence the satisfiability threshold on the balanced random formulas. All our

doi:10.1088/1742-5468/2011/03/P03023
Figure 4. Left: $\mathcal{L}_N(s)$ versus $s$ for regular random graphs with random negations, computed exactly for $15 \leq N \leq 60$ with $N_s = 10^5$ samples $\mathcal{F}$, and binning interval $\Delta s = (s_{\text{max}} - s_{\text{min}})/100$ with $K = 3, L = 8$, where $s_{\text{max}}$ and $s_{\text{min}}$ are the maximum and minimum entropies of the $N_s$ samples respectively. We also plot $\mathcal{L}(s)$ versus $s$ computed with the population dynamics for both negative $x$ ($P = 5 \times 10^4$) and positive $x$ ($P = 2 \times 10^4$). Right: $\mathcal{L}_N(s)$ versus $s$ for regular random graphs with balanced negations, computed exactly for $39 \leq N \leq 87, N_s = 10^5$, binning interval $\Delta s = (s_{\text{max}} - s_{\text{min}})/100$, and $K = 3, L = 8$. The curves do not superpose, so the large deviations decay faster than exponentially.

observations about the large deviation function suggest that the threshold for the random adversarial satisfiability problem satisfies $\alpha_a > 3.39$.

6. Numerical results for AdSAT and large deviations

In this final section we compare theoretical predictions from the cavity method with numerical results.

6.1. Numerical results for large deviations

First, we investigate numerically the number of configurations of negations yielding a formula with a certain entropy of solutions. For one given random graph geometry of size $N$, we generate independently at random $I \gg 1$ different configurations of negations, and for each of them we count the number of solutions using a publicly available implementation of the exact counting algorithm relsat [29]. We define the probability $P_N(s)$ over the negation-configurations that the value of the entropy density is between $s$ and $s + \Delta s$, where $\Delta s$ is a binning interval that we will specify later.

Following the assumption of exponentially-small large deviations made in equation (4), we define

$$\mathcal{L}_N(s) = \frac{1}{N} \left[ \log P_N(s) - \log \max_s P_N(s) \right].$$

The numerical result for $\mathcal{L}_N(s)$ is depicted in figure 4 (left) for $L = 8$, and compared to the predictions of the large deviations of the Bethe entropy from section 5. The agreement between the numerical data point and the theoretical prediction is not good in the low
Adversariable satisfiability problem

entropy region. One possibility is that this is due to pre-asymptotic effects; on the other hand this does not seem likely as the numerical curves seem to superpose nicely for different system sizes. Another possibility is that we neglected some replica symmetry breaking effects; note, however, that the large deviation calculation over survey propagation did not provide any non-trivial result. We hence leave this disagreement as an open problem.

At this point we want to recall the result from BP that we obtained on balanced regular instances with even degree (i.e. for instance $L = 8$); in that case the BP fixed point was factorized and independent of the negation-configuration even for small graphs. Let us hence investigate the numerical results for the large deviations of the entropy in this case. The data for $\mathcal{L}_N(s)$ are depicted in figure 4 (right); recall from table 1 that the maximum of the curve corresponds to $s_B$ obtained with BP. The curves in figure 4 (right) clearly do not superpose for different system sizes. Indeed, $\mathcal{L}_N(s)$ seems to be 'closing'. From these data it is indeed plausible that in the limit $N \to \infty$, $\mathcal{L}_N(s)$ converges to a delta function on the value of entropy $s = s_B$.

Hence, the data in figure 4 (right) suggest that the probability that the entropy of a formula is different from the value predicted by BP is smaller than exponentially small. This makes us conclude that in a general case, the probability of having an entropy outside the interval $(s_L, s_R)$ is smaller than exponentially small (we recall that for the balanced negations and even degree $L$ regular graphs $s_L = s_R = s_B$). Hence in the thermodynamic limit there are almost surely no negation-configurations that would lead to a value of entropy outside the range $(s_L, s_R)$.

Moreover, the large deviation function $\mathcal{L}_N(s)$ if asymptotically negative can be interpreted as a probability of generating a rare graph and configuration of negations having entropy $s$ [30]. Since there are of order $N^N$ regular graphs, and there is no or at least one graph with entropy $s \notin (s_L, s_R)$, we can have either $P_N(s) = 0$ or

$$ P_N(s) \geq e^{-c_1 N \log N}, \quad (28) $$

where $c_1$ is some positive constant. Consider now that there are $e^{N \mathcal{L}'}$ of configurations of negations (e.g. $\mathcal{L}' = K \alpha \log 2$ if we consider all the negation-configurations, or $\mathcal{L}' = \mathcal{L}_B$ if we consider just the balanced negation-configurations). The fraction of graphs with configurations of negations leading to entropy $s \notin (s_L, s_R)$ has to be small only if

$$ P_N(s) e^{N \mathcal{L}'} \ll 1. \quad (29) $$

If an equality holds in equation (28) then equation (29) holds in the thermodynamic limit, $N \to \infty$. However, (29) does not have to hold for finite $N$ unless

$$ N \geq N_c \equiv \exp \left( \frac{\mathcal{L}'}{c_1} \right). \quad (30) $$

Since $c_1$ can be considerably smaller than $\mathcal{L}'$, the crossover value of $N_c$ might be very large and out of reach for exact numerical methods. This justifies the presence of strong pre-asymptotic effects for the system sizes treated in figure 4.

6.2. Strong finite size corrections for the AdSAT threshold

We investigate numerically the AdSAT threshold $\alpha_a$ by computing the probability (over random graph instances) $p_s$ that an adversary is not able to find a configuration of
negations that makes the formula unsatisfiable. We do this on regular instances because
of the reduced fluctuations that arise due to the randomness of the graph.

We generate $I \gg 1$ regular instances for each value of the degree $L$ and for each
size $N$. Then for each instance we use simulated annealing on the negation-variables in
order to minimize the number of solutions; we monitor whether an unsatisfiable formula
is generated or not. This general strategy for AdSAT was suggested by [31]. In particular,
we introduce an inverse temperature $\beta$. Initially we set $\beta = 1$. We choose randomly one
of the negation-variables, $J_{ia}$, and attempt to flip it, i.e. to set $J_{ia} \rightarrow 1 - J_{ia}$. Denoting
by $J'$ the configuration of negations after this flip, we accept the flip with probability
$\min\{1, e^{-\beta(S_J' - S_J)}\}$. The entropy $S_J$ is computed exactly with a publicly available
implementation of the exact exhaustive search algorithm relsat [29]. This algorithm has
an exponential running time in the size of the system, limiting us to very small system
sizes. Attempting for $N$ negation flips is one Monte Carlo (MC) step. Every 10 MC
steps we multiply the inverse temperature by a rate factor $r > 1$. We keep track of the
so far minimal value of entropy $s_{\text{min}}$ and the index $n_0$ of the MC step in which it was
first found. The algorithm stops if either an unsatisfiable instance is encountered or no
further decrease in the value of entropy $s_{\text{min}}$ has occurred in the last $9 \times n_0 + 50$ MC steps.
The probability $p_s$ plotted in figure 5 is then given by the fraction of cases in which an
unsatisfiable instance was not found.

There is of course no guarantee that our algorithm found the actual minimal possible
entropy. So, strictly speaking, any result for the satisfiability threshold derived from the
data for $p_s$ is only an upper bound to the true threshold. However, given the strictness of
our stopping condition we have a reasonable confidence that our results are very close to
the exact results. Figure 5 depicts the fraction of regular instances of size $N$ where we were
unable to find a configuration of negations that would make the formula unsatisfiable.

On first sight the numerical data in figure 5 do not agree with our theoretical
predictions. Indeed, we predicted that unsatisfiable configurations of negations exist only
for $L \geq 11$, whereas, for the system sizes our simulated annealing algorithm was able

Figure 5. Probability $p_s$ that using the simulated annealing algorithm described
in the text we did not find any unsatisfiable configuration of negations for $L$
regular instances, as a function of $L$, for different system sizes $9 \leq N \leq 54$. We
used the annealing rate $r = 1.1$, and number of instances $I = 100$. 
to treat, we find unsatisfiable negation-configurations for a large fraction of graphs with $L \geq 8$.

On a second sight, however, we see in figure 5 that for $L = 6$ and 7 there are very strong finite size corrections to $p_s$. Indeed, for $L = 6$ and size $N = 9$ we find that roughly $3/4$ of the instances can be made unsatisfiable, whereas for $N = 36$ none of the $I = 100$ instances that we tried can be made unsatisfiable. Similarly for $L = 7$ and size $N = 36$ we find that most of the $I = 100$ instances can be made unsatisfiable, whereas for $N = 54$ almost none of them. If this trend continues it is perfectly plausible that in the $N \to \infty$ limit even for $L = 10$ the adversary is never successful. These results, in agreement with the conclusions of section 6.1, suggest very strong pre-asymptotic effects in the AdSAT problem. The strength of the finite size corrections hence poses a challenge to numerical verifications of our cavity method asymptotic predictions.

On the other hand, the scaling argument presented in equation (30) suggests that the system sizes at which the asymptotic behavior starts to be dominant might be quite large (perhaps thousands or more); this is in particular true in the vicinity of the satisfiability threshold. Hence, in the AdSAT problem, and likely also in other adversarial optimization problems, it is particularly important to develop techniques that predict the pre-asymptotic behavior and the finite size corrections. We saw from the results on random regular instances with even degree that BP predicts the same Bethe entropy for all balanced negation-configurations independently of the system size; hence the methods for analysis of finite size corrections and pre-asymptotic effects will have to go beyond the assumptions of the cavity method. On the other hand, analysis of the cases where the BP fixed point is factorized might be a good playground for the development of such techniques.

7. Discussion and conclusions

In this paper we studied the adversarial satisfiability problem and concluded that the most frustrated instances of random $K$-SAT are very close to the ones with balanced configurations of negations. For random regular 3-SAT instances this leads to a threshold $L = 11$, starting from which the adversary is able to find unsatisfiable configurations of negations (compare to $L = 14$ for the ordinary random regular 3-SAT). For the canonical (Poissonian) adversarial 3-SAT this leads to $\alpha_a = 3.399(1)$ (compare to $\alpha_c = 4.2667$ for the ordinary random 3-SAT). The satisfiability threshold values for the regular and for the balanced 3-SAT instances were obtained as a side result.

This result is rather uninteresting from the algorithmic point of view, as balancing negations is an easy problem. However, the same method we used here can be applied to more interesting situations, for instance the quantified SAT problem. Recall also that the adversarial satisfiability problem was suggested as a problem interpolation between random SAT and random quantum SAT. Note, however, that our study leads to the conclusion that the adversary SAT is much closer to the classical random SAT than to the quantum SAT. Note that in the large $K$ limit the random satisfiability threshold scales as $\alpha \approx 2^K \log 2$; the same scaling holds for the threshold in the random $K$-SAT with balanced negations, since at large $K$ the degree of the variables is so large that the difference between the Poissonian distribution of the number of non-negated variables and balanced negation-configurations does not play any role in the leading order in $K$. On
the other hand, the satisfiability threshold of the quantum SAT was upper-bounded by $2^K \log 2/2$ \cite{17,18}; hence the quantum effect must be responsible for this drastic decrease of the threshold value.

We obtained our results by studying the large deviations of the entropy in the ordinary random $K$-SAT. In particular, an approach leading to equations very similar to the 1RSB equations leads to the calculation of the large deviations in the case where rare instances are exponentially rare. Exponential large deviations are common in statistical physics. In some cases, see, e.g., \cite{32}, \cite{11}–\cite{13}, the large deviations are rarer than exponentially rare. In our study this arises for regular random $K$-SAT instances with balanced negations and even degree. In cases where the large deviation function decays faster than exponentially with the system size extremely strong finite size corrections and pre-asymptotic effects can be induced, as we argued in section 6, where we presented numerical studies of the large deviations and of the satisfiability threshold. Interestingly, methods based on the standard cavity method are not straightforwardly applicable to study of the related finite size corrections and pre-asymptotic behavior. It remains a theoretical challenge to find out how to describe analytically and algorithmically these pre-asymptotic effects that might be crucial for solving some industrial instances of adversarial optimization problems.

Finally we did not address the possibility of replica symmetry breaking in the space of negations (due to its technical difficulty) and this should also be a subject of future works.

Acknowledgments

We thank Cris Moore for introducing us to the adversarial SAT problem, Antonello Scardicchio for very helpful discussions, and Guilhem Semerjian for very helpful discussions and very useful comments about a preliminary version of the paper. We also acknowledge support from the DI computational center of University Paris-Sud.

References

[1] McMasters A W and Mustin T M, Optimal interdiction of a supply network, 1970 Nav. Res. Logist. Q. 17 261
[2] Wood R K, Deterministic network interdiction, 1993 Math. Comput. Modelling 17 1
[3] Varga P, Minimax games, spin glasses, and the polynomial-time hierarchy of complexity classes, 1998 Phys. Rev. E 57 6487
[4] Papadimitriou C H, 1994 Computational Complexity (Reading, MA: Addison-Wesley)
[5] Moore C and Mertens S, 2011 The Nature of Computation (Oxford: Oxford University Press)
[6] Cook S A, The complexity of theorem-proving procedures, 1971 Proc. 3rd STOC (New York, NY) ACM pp 151–8
[7] Mitchell D G, Selman B and Levesque H J, Hard and easy distributions for SAT problems, 1992 Proc. 10th AAAI (Menlo Park, CA: AAAI Press) pp 459–65
[8] M´ezard M, Parisi G and Zecchina R, Analytic and algorithmic solution of random satisfiability problems, 2002 Science 297 812
[9] Krzakala F, Montanari A, Ricci-Tersenghi F, Semerjian G and Zdeborová L, Gibbs states and the set of solutions of random constraint satisfaction problems, 2007 Proc. Nat. Acad. Sci. 104 10318
[10] Rivoire O, The cavity method for large deviations, 2005 J. Stat. Mech. P07004
[11] Parisi G and Rizzo T, Large deviations in the free energy of mean-field spin glasses, 2008 Phys. Rev. Lett. 101 117205
[12] Parisi G and Rizzo T, Phase diagram and large deviations in the free energy of mean-field spin glasses, 2009 Phys. Rev. B 79 134205
[13] Parisi G and Rizzo T, Large deviations of the free energy in diluted mean-field spin-glass, 2010 J. Phys. A: Math. Theor. 43 045001

doi:10.1088/1742-5468/2011/03/P03023
Adversarial satisfiability problem

[14] Parisi G, A sequence of approximated solutions to the SK model for spin-glasses, 1980 J. Phys. A: Math. Gen. 13 L115
[15] Dotsenko V, Franz S and Mezard M, Partial annealing and overfrustration in disordered systems, 1994 J. Phys. A: Math. Gen. 27 2351
[16] Laumann C R, Moessner R, Scardicchio A and Sondhi S L, Phase transitions and random quantum satisfiability, 2010 Quantum Inf. Comput. 10 1
[17] Bravyi S, Moore C and Russell A, Bounds on the quantum satisfiability threshold, 2009 arXiv:0907.1297v2
[18] Laumann C R, Lauchli A M, Moessner R, Scardicchio A and Sondhi S L, Product, generic, and random generic quantum satisfiability, 2010 Phys. Rev. A 81 062345
[19] Chen H and Interian Y, A model for generating random quantified boolean formulas, 2005 IJCAI 2005: Proc. 19th Int. Joint Conf. on Artificial Intelligence pp 66–71
[20] Altarelli F, Braunstein A, Ramezanpour A and Zecchina R, Statistical physics of optimization under uncertainty, 2010 arXiv:1003.6124v1, 2010
[21] Mertens S, M´ezard M and Zecchina R, Threshold values of random k-SAT from the cavity method, 2006 Random Struct. Algorithms 28 340
[22] M´ezard M and Parisi G, The Bethe lattice spin glass revisited, 2001 Eur. Phys. J. B 20 217
[23] Yedidia J S, Freeman W T and Weiss Y, Understanding belief propagation and its generalizations, 2003 Exploring Artificial Intelligence in the New Millennium (San Francisco, CA: Morgan Kaufmann) pp 239–6
[24] M´ezard M and Montanari A, 2009 Physics, Information, Computation (Oxford: Oxford Press)
[25] Montanari A, Parisi G and Ricci-Tersenghi F, Instability of one-step replica-symmetry-broken phase in satisfiability problems, 2004 J. Phys. A: Math. Gen. 37 2073
[26] M´ezard M and Zecchina R, Random k-satisfiability problem: from an analytic solution to an efficient algorithm, 2002 Phys. Rev. E 66 056126
[27] Zdeborov´a L and Krzakala F, Phase transitions in the coloring of random graphs, 2007 Phys. Rev. E 76 031131
[28] Montanari A, Ricci-Tersenghi F and Semerjian G, Clusters of solutions and replica symmetry breaking in random k-satisfiability, 2008 J. Stat. Mech. P04004
[29] Bayardo R J Jr and Pehoushek J D, Counting models using connected components, 2000 Proc. 17th AAAI (Menlo Park, CA: AAAI Press) pp 157–62
[30] Rivoire O, Properties of atypical graphs from negative complexities, 2004 J. Stat. Phys. 117 453
[31] Nagaj D and Scardicchio A, 2010 in preparation
[32] Monthus C and Garel T, Matching between typical fluctuations and large deviations in disordered systems: application to the statistics of the ground state energy in the sk spin-glass model, 2010 J. Stat. Mech. P02023