Source coding by efficient selection of ground-state clusters

PHYSICAL REVIEW E **72**, 015103(R) (2005)

# Source coding by efficient selection of ground-state clusters

Demian Battaglia,[1] Alfredo Braunstein,[1,2] Joël Chavas,[2] and Riccardo Zecchina[2]
[1]*SISSA, Via Beirut 9, I-34100 Trieste, Italy*
[2]*ICTP, Strada Costiera 11, I-34100 Trieste, Italy*
(Received 24 December 2004; published 18 July 2005)

We analyze the geometrical structure of clusters of ground states which appear in many frustrated systems over random graphs. Focusing on the regime of connectivities where the number of clusters is exponential in the size of the problems, we identify an appropriate generalization of the survey propagation equations efficiently exploring the geometry. The possibility of selecting different clusters has also computational consequences. As a proof of concept here we show how a well-known physical system can be used to perform nontrivial data compression, for which we introduce a unique compression scheme. Performances are optimized when the number of well-separated clusters is maximal in the underlying physical model.

DOI: 10.1103/PhysRevE.72.015103 PACS number(s): 02.50.−r, 05.20.−y, 75.10.Nr, 89.70.+c

The combinatorial problem of satisfying a large set of constraints that depend on $N$ discrete variables is a fundamental one in statistical physics of disordered (frustrated) systems as well as in computer science and engineering [1]. Even for randomly generated problem instances, asking whether it exists an assignment to the variables that satisfies all constraints simultaneously seems to become extraordinarily difficult to solve as some control parameters are varied. Recent advances in the statistical-mechanics study of random constraint satisfaction problems (CSPs) have identified the origin of such difficulty in a dynamical spin-glass transition: inside the glassy phase, the space of optimal configurations becomes divided into an exponential number of clusters. Associated with this equilibrium set of states, there exists an even larger set of metastable states [2–5], which act as dynamical traps for local dynamical processes.

Similar features are found in a large variety of models, when defined on random graphs, ranging from physical systems (spin glasses, kinetically constrained models, rigidity percolation) to computer science problems (combinatorial optimization), to engineering problems (error correcting codes, control networks).

An important by-product of the analytical studies of random CSPs has been the introduction of a new class of algorithms—the so-called survey propagation (SP) algorithms [5–7]—specially devised to deal with the clustering scenario and able to find optimal assignments of benchmark problems on which all other known optimization algorithms fail.

In this Rapid Communication we make a step forward in the understanding of the potentialities of the statistical physics approach, by addressing the following questions:

(i) Given a frustrated system defined over a random structure (graph), is it possible to explore efficiently its ground states space?

(ii) Can we use the capability of addressing a large set of states for computational purposes?

The first question relates to the general physical issue of probing the topology of the space of solutions (ground states) in problems that are in a clustered phase, the so-called replica symmetry-breaking (RSB) phase. Surprisingly enough, such geometrical insight is also important for engineering applications in error correcting codes [8]. The second question addresses a new algorithmic perspective in which the presence of many states becomes a resource for a computational device.

In what follows we shall provide a positive answer to both questions by identifying a generalization of the SP equations which are indeed capable of addressing efficiently—with a computational cost almost linear in the size of the problems—an exponential number of different clusters of ground states. From the physical side, we provide exact numerical evidence for the RSB geometric structure for large size instances of diluted spin glasses and optimization problems (similar results had been obtained only in limit-case systems for which the problem of finding ground states is tractable in full generality [9]). On the computational side, we take advantage of the addressability of the set of clusters of ground states to produce a "physical" lossy data compression scheme [10].

A generic constraint satisfaction problem is defined by $N$ discrete variables which interact through constraints involving typically a small number of variables. The energies $C_a(\vec{\xi})$ of the single constraints (equal to 0 or 1, depending on if they are satisfied or not by a given assignment $\vec{\xi}$) sum up to give the global energy function $\mathcal{E}$ of the problem and are function of just a small subset of variables $V(a) = \{j_{a_1}, \ldots, j_{a_K}\}$ [every variable $j$ is involved, on the other hand, in a subset $V(j)$ of constraints].

Since one is interested in satisfying all the constraints simultaneously, the CSP is just equivalent to the problem of looking for zero-energy ground states. For most NP-complete CSP the function $\mathcal{E}$ can be directly interpreted as a spin-glass-like Hamiltonian [3,6,11]. For instance, the well-known case of the random $K$-satisfiability ($K$-SAT) problem consists of deciding if $M$ clauses—taking the form of the OR function of $K$ variables chosen randomly among $N$ possible ones—can be simultaneously true. The energy contribution associated with a single clause can then be written as $C_a(\vec{\xi}) = \prod_{l=1}^{K} \frac{1}{2}(1 + J_{a,l} x_{a_l})$, where $x_{a_l} = \pm 1$ depending on the truth-value of $j_{a_l}$ and $J_{a,l} = \pm 1$ if $j_{a_l}$ appears negated or directed in the clause $a$. The same variable can appear directed and negated in different terms and hence give rise potentially to frustration.

In the so-called factor graph representation, variables and constraints are represented by nodes, with links connecting a variable node with a constraint one if the latter depends functionally on the former. The connectivity distribution and the loop structure of the *factor graph* associated with a CSP have a strong influence on the behavior of search algorithms. For many important random CSPs, when the ratio $\alpha=M/N$ is included in a narrow region $\alpha_d<\alpha<\alpha_c$—the exact values of the thresholds depending on the details of the problem and on the disorder—the problem is still satisfiable but the zero-energy phase of the associated Hamiltonian breaks down in an exponential number of clustered components. The cavity method provides accurate analytical computations [6,12,13] of the thresholds' location in good agreement with the numerical experiment [14], and provide as well the theoretical foundation of the survey propagation message passing algorithm, successful in the resolution of instances of both the $q$-coloring and the $K$-SAT problems [6,11], which are hard for local search algorithms. A constraint node $b$ is supposed to send a message $\vec{u}_{b\to j}=\vec{e}_s$ (where $\vec{e}_s$ are vectors having just the $s$th component equal to 1) to a variable $j$ each time that $j$ would violate the constraint $b$ assuming the value $s$. For locally treelike factor graphs (like the ones associated typically with randomly generated instances) the messages incoming to $j\in V(a)\backslash i$ can be assumed uncorrelated, after the temporary removal of a single clause $a$ and of a variable $i$ (the so-called cavity step). It becomes then possible to evaluate probability density functions for the messages, called *cavity surveys* (the probability space originates from the set of all clusters of satisfying assignments sampled with uniform measure),

$$Q_{b,j}(\vec{u}_{b,j}) = \eta^0_{b,j}\delta(\vec{u}_{b,j},\vec{0}) + \sum_{s=1}^{q}\eta^s_{b,j}\delta(\vec{u}_{b,j},\vec{e}_s). \qquad (1)$$

Here, $\eta^s_{b,j}$ are the probabilities that $b$ constrains $j$ not to enter the $s$ state and $\eta^0_{b,j}$ is the probability that no bias is induced, $b$ being already satisfied by the assignment of other variables.

The cavity surveys form a closed system of $KM$ functional equations for which a solution can be found in linear time by iteration [6,7],

$$Q_{a,i}(\vec{u}_{a,i}) = \int \mathcal{DQ}\ \delta_{\mathcal{E}}[\{\vec{u}_{b,j}\}]\chi(\vec{u}_{a,i},\{\vec{u}_{b,j}\}), \qquad (2)$$

where the function $\chi(\{\vec{u}_{b,j}\})$ depends on the specific CSP and $\mathcal{DQ}=\Pi_{j\in V(a)\backslash i}\Pi_{b\in V(j)\backslash a}Q_{b,j}(\vec{u}_{b,j})$. The functional $\delta_{\mathcal{E}}[\{\vec{u}_{b,j}\}]$ acts as a filter, assigning null weight to sets of messages associated with clusters of excited configurations. At the fixed point, from the knowledge of the surveys one may compute the fractions $W^s_j(W^0_j)$ of clusters of solutions in which a variable $j$ is frozen in the direction $s$ (or is unfrozen). Such microscopic information can be successfully used to find optimal assignments by *decimation* [7].

We shall now present a generalization of the SP algorithm (SP-ext), allowing the retrieval of a solution close to any desired configuration $\vec{\xi}$. Hereafter we shall refer for simplicity to the $K$-SAT problem ($s=\pm1$ only), but the method could be easily extended to a generic CSP. On general grounds, a way to analyze specific regions of the configuration space would be to solve the cavity equations in the presence of an additional field conjugated to some geometrical constraint (e.g., fixed magnetization). However, this would lead to an algorithmically inefficient scheme. We consider instead an arbitrary but quite natural extension of the SP equations in which external messages $\vec{u}_i=\vec{e}_{-\xi_i}$ in an arbitrary direction $\vec{\xi}\in\{-1,1\}^N$ are introduced for each variable. New associated surveys $Q^\pi_i(\vec{u}_i)=(1-\pi)\delta(\vec{u}_i,\vec{0})+\pi\delta(\vec{u}_i,\vec{e}_{-\xi_i})$ are given *a priori* and never updated, and affect dynamically the relative weight of the different clusters, entering into the measure $\mathcal{DQ}$ in the convolution integrals (2). The parameter $\pi$ can be interpreted as strength of the perturbation. Convergence can be reached only if the zero-energy constraint is respected. While an intensity $\pi\simeq1$ would produce a complete polarization of the messages if $\vec{\xi}$ was a solution, in the general case the use of a smaller forcing intensity allows the system to react to the contradictory driving and to converge to a set of surveys sufficiently biased in the desired direction, allowing for an efficient selective exploration of specific parts of the solution space.

In order to use SP-ext for probing the local geometry of the zero-energy phase one proceeds as follows. First, a random solution $\vec{\sigma}$ is found by decimation. Next, new satisfying assignments $\vec{\sigma}_d$ are generated, by forcing the system along a direction obtained flipping $Nd$ spins of the original solution $\sigma$. In order to have a highly homogeneous distribution of the clusters, we have chosen for our experiments an ensemble of random $K$-SAT in which variables have fixed degrees and are balanced (i.e., have an equal number of directed and negated occurrences in the clauses). For this specific ensemble and for $K=5$, one has $\alpha_d=14.8$ and $\alpha_c=19.53$. Between $\alpha_d$ and $\alpha_G=16.77$ the phase is expected to be unstable, whereas between $\alpha_G$ and $\alpha_c$ the 1-RSB phase is stable. We have estimated $\alpha_G$ with a new message passing algorithm implementing the cavity equations at the level of two steps of RSB [15].

In the experiments we have taken instances of size $N=10^4$ with an intensity of the forcing $\pi=0.35$ (the maximal value allowing for convergence). The Hamming distance between $\sigma$ and $\zeta_d$ is plotted against $d$ in the first *stability diagram* (black data points) shown in Fig. 1. For $d<d_c\simeq0.3$, $D(\vec{\sigma},\vec{\zeta}_d)$ linearly increases with a very small slope until a value $d_{\rm cl}\simeq0.1$. Conversely, for $d>d_c$, it jumps to a value $d_0-\Delta\simeq0.28$, and a symmetric distribution of distances around $d_0$ is obtained. Under the hypothesis of homogeneous distribution of clusters, the fixed-point average site magnetization $\langle W^+-W^-\rangle$ provides an analytic estimation of the typical overlap $q_0=1-d_0/2$ between two different clusters in agreement with the experiments. On the other hand, $d_{\rm cl}$ is of the order of the average fraction of unfrozen variables $\langle W^0\rangle$. The gap between clusters is the main prediction of the 1-RSB cavity theory which is nicely confirmed by these experiments.

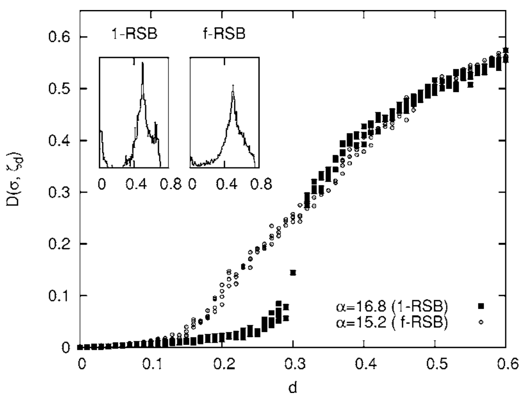


FIG. 1. Distribution of distances. A solution $\zeta_d$ is generated by forcing along a vector with $d$ spins flipped with respect to a reference solution $\sigma$. The difference between the 1-RSB and $f$-RSB cases appears by both looking at the stability diagrams themselves (white circles against black squares) and at the reciprocal distances histograms.

A completely different behavior is observed when repeating the experiment in the expected 1-RSB unstable phase. The histogram of the reciprocal distances among all the generated solutions (inset of Fig. 1, $\alpha$=15.2) is now gapless and the related stability plot in the main figure (white data points) deviates significantly from the 1-RSB case.

The ability of SP-ext to select clusters is unique computational feature which may play a role in different systems in which clustering of input data is important. As a proof of concept, here we provide a first simple application by implementing a lossy data compressor [10] which exploits the 1-RSB clustered structure for data quantization purposes.

Let us suppose to have an $N$-bit binary input string $\vec{\xi}$ generated from an unbiased and uncorrelated random source. Given an appropriate $K$-SAT instance with $N$ variables, a solution $\vec{\sigma}_\xi$ as close as possible to $\vec{\xi}$ can be generated with SP-ext. One can expect to find a solution at a distance close to $d_0-\Delta$, if the cluster distribution is homogeneous (balanced and fixed even connectivity instances are then chosen). Furthermore, $\alpha$ is taken slightly larger than $\alpha_G$, in order to maximize the number of addressable clusters, still preserving a sharp separation among them. At this point, a compressed string $\vec{\sigma}_\xi^R$ is built by retaining just the spins of the first $NR$ variables of $\vec{\sigma}_\xi$. In the decompression stage, SP-ext is run over the same graph, applying a very intense forcing ($\pi$ =0.99) parallel to $\vec{\sigma}_\xi^R$. If $R>R_c$ (where $R_c$ becomes quickly independent on $N$ as $N$ increases, for fixed graph ensemble and $K$), SP-ext becomes able to select exactly the single cluster to which $\vec{\sigma}_\xi$ belongs. The cluster addressing is actually so sharp, that no decimation is needed and all the remaining $N(1-R)$ unforced variables can simultaneously be fixed to their preferred orientation without creating contradictions. A comparison with the theoretical Shannon bound [10] is done in Fig. 2 (dotted line), where the cluster selection transition is clearly visible. The accumulated distortion with respect to $\vec{\xi}$ is of the order of $d_{\rm cl}+d_0-\Delta$. The lines relative to the performance of a trivial decoder in which the missing bits are randomly guessed and of a standard textbook code, the repetition code, are also plotted for comparison.

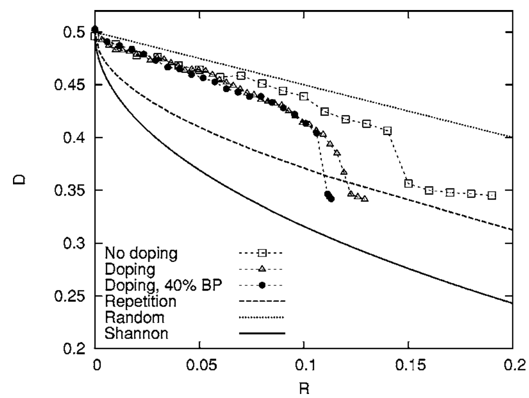


FIG. 2. Rate-distortion profile for the compression of a random source. Various decompression methods are used—simple cluster reconstruction, reconstruction with doping, and with doping and SP/BP interpolation. The performance of repetition codes and of random guessing is also plotted for comparison.

Better compression can be achieved by the use of the *iterative doping* [16] technique for choosing the bits to store. After the determination of $\vec{\sigma}_\xi$, SP-ext is run again *without* applying any forcing and a ranking of the most balanced variables is performed. One looks for the variable $i$ which minimizes $|W_i^+ - W_i^-| + W_i^0$ (frozen in opposite directions in a similar number of clusters and rarely unconstrained). The state assumed by $i$ in the solution $\vec{\sigma}_\xi$ will be taken as the first bit of the compressed string $\vec{\sigma}_\xi^R$ and used to fix $i$. New doping steps are done until the desired compression rate has been reached. An *identical* doping stage is then performed in decompression. The iterative ranking allows indeed to find out which spins have to be fixed accordingly to the ordered bit sequence $\vec{\sigma}_\xi^R$ and the left variables can be fixed as in the previous decompression method. Fixing a balanced variable "switches off" a larger number of clusters, and, hence, the critical rate is reduced (dash-dotted curve in Fig. 2), thanks to a less redundant coding of the information needed for the cluster selection. A further improvement can be obtained by using in the doping stage a modified iteration that interpo-

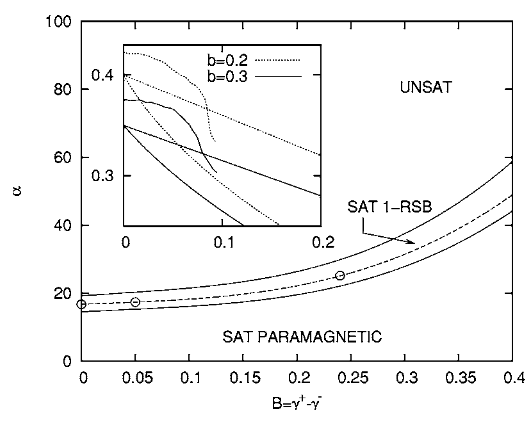


FIG. 3. Phase diagram for fixed degree 5-SAT and compression performance for two biased sources ($b$=0.2 and $b$=0.3). The white circles—following closely the Gardner instability line—show the position of the best instances found for the compression of differently biased sources.

lates between SP and the well-known belief propagation equations (solid curve in Fig. 2), which carry information about the intracluster bias of the variables [17].

We note that a careful analysis of finite-size effects, sampling sizes from $N \sim O(10^2)$ to $N \sim O(10^5)$, confirms the stability of both the gap and the cluster retrieval capabilities of the scheme.

SP-ext can also be applied when the occurrence probability $P_{\pm}$ of the possible input symbols are different. The issue is of practical relevance since correlated sources can often be shown to be equivalent to memoryless biased sources [18]. Let us suppose that the bias of the source is $b = P_+ - P_- > 0$. It is possible to engineer graphs with a cluster distribution concentrated around the ferromagnetic direction, by making for every variable $i$ the fraction $\gamma_+$ of couplings $J_{a,i} = +1$ larger than the fraction $\gamma_-$ of $J_{a,i} = -1$. As shown indeed in Fig. 3, a narrow SAT RSB stripe is still present for a balancing $B = \gamma_+ - \gamma_- < 0.435$. When the balancing is too large, there is, on the other hand, a direct transition between an unfrozen SAT phase and the UNSAT region. The rate-distortion profile of the compression of a random uncorrelated source with $b = 0.2$ and $b = 0.3$ is shown in the inset of the same figure. The best graphs found for all the analyzed values of $b$ are always located in proximity of the Gardner line (empty circles in Fig. 3).

It is expected that better performances can be achieved by a careful optimization of the graph ensemble as it was shown to be the case for iterative decoding with belief propagation in error correcting codes [19].

This work has been supported by STIPCO and EVERGROW EU projects.

[1] M. R. Garey, D. S. Johnson, *Computers and Intractability* (Freeman, New York, 1979).
[2] R. Monasson, R. Zecchina, S. Kirkpatrick, B. Selman, and L. Troyanski, Nature (London) **400**, 133 (1999).
[3] NP-Hardness and Phase Transitions, edited by O. Dubois, R. Monasson, B. Selman, and R. Zecchina, Special Issue of Theor. Comput. Sci. **265**, Issue: 1–2 (2001).
[4] G. Biroli, R. Monasson, and M. Weigt, Eur. Phys. J. B **14**, 551 (2000).
[5] M. Mézard, G. Parisi, and R. Zecchina, Science **297**, 812 (2002).
[6] M. Mézard and R. Zecchina, Phys. Rev. E **66**, 056126 (2002).
[7] A. Braunstein, M. Mézard, and R. Zecchina, e-print cs.CC/0212002, Random Struct. Algorithms (to be published).
[8] T. Richardson and R. Urbanke, in *Codes, Systems, and Graphical Models*, edited by B. Marcus and J. Rosenthal (Springer, New York, 2001).
[9] S. Franz, M. Leone, F. Ricci-Tersenghi, and R. Zecchina, Phys. Rev. Lett. **87**, 127209 (2001).
[10] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms* (Cambridge University Press, Cambridge, MA, 2003).
[11] A. Braunstein, R. Mulet, A. Pagnani, M. Weigt, and R. Zecchina, Phys. Rev. E **68**, 036702 (2003).
[12] S. Mertens, M. Mézard, and R. Zecchina, e-print cs.CC/0309020.
[13] A. Montanari, G. Parisi, and F. Ricci-Tersenghi, J. Phys. A **37**, 2073 (2004).
[14] S. Kirkpatrick and B. Selman, Science **264**, 1297 (1994).
[15] C. Baldassi and R. Zecchina (unpublished).
[16] G. Caire, S. Shamai, S. Verdù, in *Proceedings of the IEEE Information Theory Workshop*, Paris (IEEE, New York, 2003), pp. 291–295.
[17] E. Maneva, E. Mossel, and M. Wainwright, in *Proceedings of SODA* (ACM, New York, 2005), p. 1089.
[18] M. Burrows, Tech. Rep. SRC 124 (1994).
[19] D. J. C. MacKay and R. M. Neal, IEEE Electron Device Lett., **32**, 1645 (1996); S. Y. Chung, G. D. Forney, T. J. Richardson, and R. Urbanke, IEEE Commun. Lett. **5**, 58 (2001).