PHYSICAL REVIEW E, VOLUME 63, 020106(R)

Entropy-based analysis of the number partitioning problem

A. R. Lima[1,2,*] and M. Argollo de Menezes[2,†]
[1]*Laboratoire de Physique et Méchanique des Milieux Hétérogènes, ESPCI Paris, 10 rue Vauquelin, 75231 Paris Cedex 05, France*
[2]*Instituto de Física, Universidade Federal Fluminense, Avenida Litorânea 24210-340, Niterói, RJ, Brazil*
(Received 27 July 2000; published 26 January 2001)

In this paper we apply the multicanonical method of statistical physics on the number partitioning problem (NPP). This problem is a basic NP-hard problem from computer science, and can be formulated as a spin-glass problem. We compute the spectral degeneracy, which gives us information about the number of solutions for a given cost E and cardinality difference m. We also study an extension of this problem for Q partitions. We show that a fundamental difference on the spectral degeneracy of the generalized ($Q>2$) NPP exists, which could explain why it is so difficult to find good solutions for this case.

DOI: 10.1103/PhysRevE.63.020106 PACS number(s): 05.20.−y, 89.20.Ff, 02.70.Rr, 75.10.Nr

I. INTRODUCTION

The use of statistical mechanics tools to understand the main ideas underlying problems as diverse as biological, social, and economic systems has become a common task, both theoretically and computationally [1,2]. Recently, these tools have been applied to computer science problems [3–11], not intending to solve them exactly, but rather to understand their complexity and the underlying mechanisms generating such complex behavior. In this Rapid Communication we focus on the number partitioning problem, a fundamental problem in theoretical computer science [12]. Our aim is to apply the multicanonical method (MUCA) [13] of statistical physics to this problem and study the behavior of nearly optimal solutions. This information is relevant for the development of new algorithms which try to find optimal solutions.

In the next section we discuss the number partitioning problem and its formulation as a spin-glass problem. We introduce the multipartitioning problem and map it onto a Q-states Potts model. In Sec. III we present the multicanonical method and apply it to our problem. In Sec. IV we discuss our results both for the classical and for the multipartitioning problem.

II. NUMBER PARTITIONING PROBLEM

The number partitioning problem (NPP) is, according to Garey and Johnson [12], one of the six basic computer science problems. Given a set $A=\{a_1,a_2,a_3,a_4,\ldots,a_N\}$ with N integer numbers, the traditional NPP consists of partitioning the set A into two disjoint sets A_1 and A_2 such that the difference

$$E=\left|\sum_{a_i\in A_1} a_i-\sum_{a_i\in A_2} a_i\right| \tag{1}$$

is minimized. If there are N_1 numbers in the set A_1 and N_2 numbers in the set A_2, then

$$m=|N_1-N_2| \tag{2}$$

is called the *cardinality difference* of the set.

On the unbalanced NPP the only condition is to minimize the cost function [Eq. (1)] without any restriction to the value of m. The problem of finding good solutions (whenever they exist) for the unbalanced (nonfixed m) NPP was essentially solved by the deterministic Karmarkar-Karp-Korf complete algorithm [14,15]. This algorithm was generalized by Mertens for the balanced (m fixed) case [16]. Some recent papers addressed the possibility of carrying a statistical analysis of this problem [5,6,8–11], obtaining interesting results, such as: the existence of an easy-to-hard transition (explained below) [5], the non-self-averaging property of the ground state energy [6], the analytical derivation of the lower bounds for the energy as a function of the cardinality difference [8,11], and the equivalence of the NPP to a random cost problem [10]. Certainly one of the most interesting features of the NPP is the existence of an easy-to-hard transition. For N independent and identically distributed (i.i.d.) random b-bit numbers a_i, the computational effort needed to obtain a solution grows exponentially with N for $N\lesssim b$ and polynomially for $N\gg b$. In this sense, there is a ''phase transition'' in the system [5,11]. For that purpose, a mapping of the NPP problem onto a spin-glass model was proposed: associate to each number a_i a new variable s_i (which we call ''spin'') such that if $a_i\in A_1$ then $s_i=-1$, otherwise $s_i=+1$. With this mapping, we can search for a configuration of spins $s_1,\ldots,s_N$, which minimizes the cost function (or energy)

$$E=\left|\sum_{i=1}^{N} s_i a_i\right| \tag{3}$$

or its square,

$$E_{SG}=\sum_{ij}^{N} J_{ij}s_i s_j, \tag{4}$$

with $J_{ij}=a_i a_j$, which we recognize as an infinite-range spin-glass Hamiltonian. We can also write the cardinality difference in a ''magnetizationlike'' way,

*Email address: arlima@if.uff.br
†Email address: marcio@if.uff.br

$$m = \frac{1}{N}\left|\sum_{i=1}^{N} s_i\right|. \tag{5}$$

Finding an optimal solution for the number partitioning problem consists of finding the spin configuration of the ground state on the spin-glass problem. This is a very difficult task, mainly because of the great number of metastable states separated by a hierarchy of increasingly high energy barriers [17].

An even more difficult problem is multipartitioning. This problem consists of partitioning the set of numbers A into Q disjoint sets. This problem has several applications [18,19], such as the division of N different jobs (computer programs) into Q computers. As far as we know, there is no theoretical study of the NPP for $Q>2$.

We can map this problem onto a Potts spin-glass by assigning to each number a_i a spin s_i that can assume integer values from 1 to Q. These spin magnitudes represent the set to which the number belongs. Hence, the energy can be written as

$$E = \sum_{i=1}^{Q}\sum_{j>i}^{Q} |\epsilon_i - \epsilon_j|, \tag{6}$$

where $\epsilon_i = \Sigma_{k=1}^{N} a_k \delta_{(s_k,i)}$ is the sum of the elements in the set i. In the same way we define the magnetization as

$$m = \frac{1}{N}\frac{\sum_{i=1}^{Q}\sum_{j>i}^{Q} |n_i - n_j|}{Q-1}, \tag{7}$$

where $n_i = \Sigma_{k=1}^{N} \delta_{(s_k,i)}$ is the number of elements in the set i. Clearly, this problem is much more complex than the traditional NPP ($Q=2$).

In the next section we show that the multicanonical method can be used to determine the spectral degeneracy of the problem, i.e., the number of solutions $g(E,m)$ that have a given energy E and magnetization m. In the statistical mechanics sense, this completely characterizes the problem, since the (dimensionless) entropy is given by $S(E,m) = \ln g(E,m)$.

III. MULTICANONICAL METHOD (ENTROPIC SAMPLING)

The multicanonical method was introduced in 1991 by Berg and Neuhaus [13], and the basic idea of this method is to sample microconfigurations of a given system by performing a biased random walk (RW) in the configuration space, which leads to another unbiased random walk (i.e., with uniform distribution) along the energy axis. This walk must have a visiting probability of each energy level E which is inversely proportional to $g(E)$, the quoted spectral degeneracy. If one can measure the transition probabilities from an energy level E to all other energy levels, one is able to obtain $g(E)$. The multicanonical method has been shown to be very efficient in obtaining satisfactory results for $g(E)$ in a large variety of problems such as evolutionary problems [20], phase equilibrium in binary lipid bilayer [21], and optimization problems [22] (for reviews of the method, see [23]). The entropic sampling method (ESM) [24], which we will use throughout this paper, has been proven to be an equivalent formulation of MUCA [25]. Here we are interested in the multiparametric formulation of the multicanonical method, since we must obtain the spectral degeneracy $g(E,m)$ [26] as a function of two parameters E and m. Let $E(X)$ and $m(X)$ be the energy and magnetization associated to the microstate X, the transition probability between two states X_i and X_f is given by

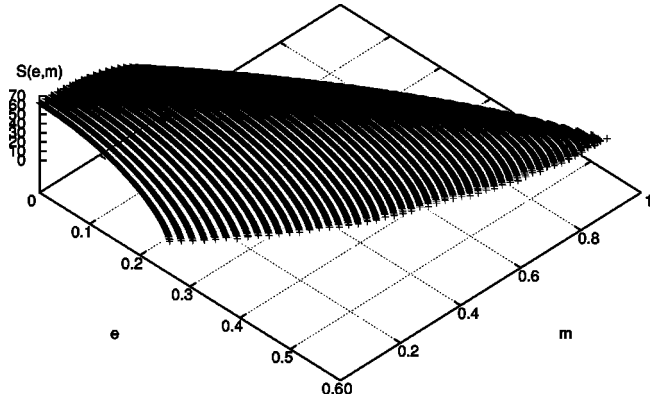

FIG. 1. Typical entropy curve as a function of cost (energy) and cardinality difference (magnetization) for the case of bipartitioning. Here the entropy in computed for a single instance of 100 numbers chosen randomly between 0 and 10^{10}. The energy is normalized by the largest possible value, $E_{\max}=10^{12}$.

$$\tau(X_i, X_f) = e^{-[S(E_f,m_f)-S(E_i,m_i)]} = \frac{g(E_i,m_i)}{g(E_f,m_f)}, \tag{8}$$

where $S(E,m)=\ln g(E,m)$ is the entropy, $E_i=E(X_i)$ [$E_f=E(X_f)$] is the energy of the initial [final] state, $m_i=m(X_i)$ [$m_f=m(X_f)$] is the magnetization of the initial (final) state, and $g(E,m)$ is the spectral degeneracy. The transitional probability [Eq. (8)] satisfies a detailed balance equation and leads to a distribution of probabilities where a state is sampled with probability $\propto 1/g(E,m)$. The successive visitations along the energy axis follow a uniform distribution, but unfortunately $g(E,m)$ is not known *a priori*. One way of obtaining $g(E,m)$ is to construct it iteratively, such as in the entropic sampling method proposed by Lee [24].

For a detailed description of the method, see Refs. [24,26,27]. We have applied this algorithm to the problem of bi- and multipartitioning in order to obtain the entropy $S(E,m)$, which is shown for the case of bipartitioning in Fig. 1. These results were obtained for a single instance (disorder realization) of 100 integer numbers chosen randomly between 0 and 10^{10}. According to the characterization of Mertens [5] for these numerical values, the problem is on the easy side of the easy-to-hard phase transition, where an exponential number of perfect solutions exist. We have performed in our simulation 2×10^7 attempts to change the state of the system [Eq. (8)]. All results shown are typical ones. Let us stress that, after obtaining the entropy of the system through extensive simulations, all thermodynamic averages

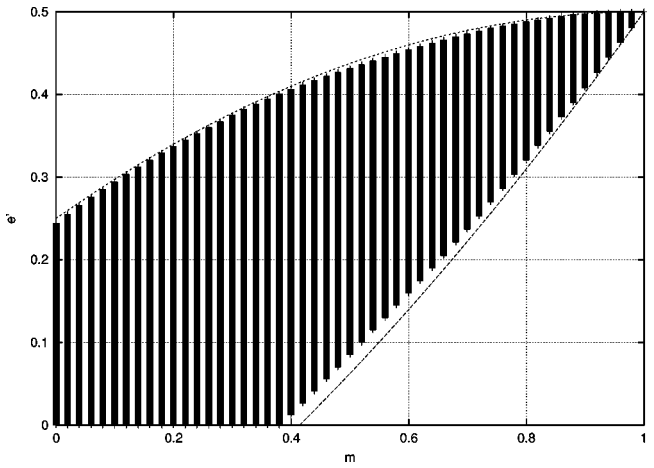

FIG. 2. Collapse of all entropies $S(e,m)$ on the (e,m) plane, evidencing the upper and lower bounds for e as a function of the cardinality difference m. Again, $e=E/E_{\max}$. The dashed lines are theoretical predictions for upper and lower bounds of the energy. Typical results for a single instance with $N=100$ numbers.

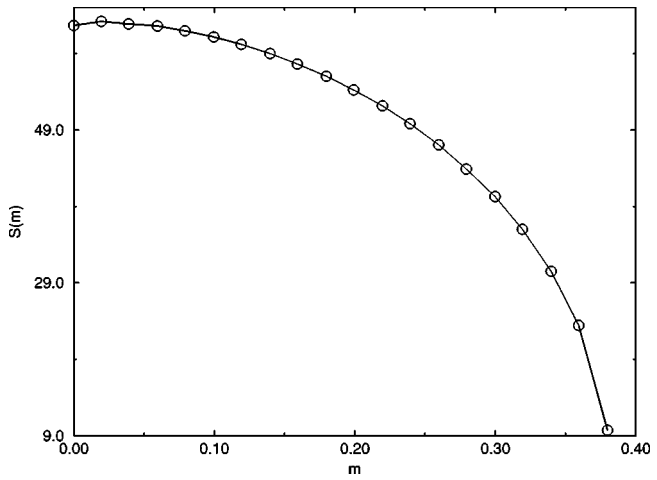

FIG. 3. Entropy of an $N=100$ instance as a function of the cardinality difference m for nearly optimal solutions. The maximum number of solutions does not correspond to an equipartition of the set, but to two subsets with $N/2-1$ and $N/2+1$ numbers, respectively. This result is not characteristic of a single instance (disorder realization).

for different cardinalities m can be obtained with no need for any further computer effort.

In the next section we are going to analyze this entropy in order to recover well-known results concerning bounds of the (E,m) curve for nearly optimal solutions. Through analysis of the entropy of the multipartitioning problem it will be clear that the change of complexity of this problem for $Q>2$ is associated with fundamental changes of the entropy curve.

IV. NUMERICAL RESULTS

Through the replica trick, Ferreira and Fontanari [8] have obtained analytical estimates for the average lower bound of the energy E as a function of magnetization (cardinality difference) m. By means of a simpler analysis, Mertens [11] has calculated lower and upper bounds for the energy $E(m)$. In Fig. 2 we show our numerical results along with analytical predictions [8,11] for both lower and upper bounds of $E(m)$. This is done by collapsing the z axis of Fig. 1 on the x,y plane [here, the (E,m) plane]. In order to compare our results with the theoretical ones, the energy is normalized appropriately such that at $m=1$, $e'=0.5$. It is important to note that our numerical result concerns only one instance, whereas the analytical results involve an average over an infinite number of instances.

In computer science one is mainly interested in optimal solutions, that is, the information contained on the first ''slice'' of the $S(E,m)$ surface, the $S(0,m)$ plane. In Fig. 3 we show $S(\epsilon,m)$, where ϵ is our numerical tolerance, which we chose to be $(E_{MAX}-E_{MIN})/1024$.

One interesting feature we have observed numerically is that the maximum number of solutions does not occur for $m=0$; it is easier to find solutions where the number of elements in each set is not exactly equal. This feature is not characteristic of a particular instance.

Now we show our results for the multipartitioning problem. If we look at the number of solutions $g(E)=\Sigma_m g(E,m)$ [or the entropy $S(E)=\ln g(E)$] for different cardinalities, we observe a fundamental difference between the $Q=2$ and $Q>2$ results (in Fig. 4 we show normalized entropies). For the $Q=2$ case, we see that the maximum of the entropy lies near $E=0$.

It is known that random movements in a statistical system leads it, on average, to the region of maximum entropy, where the number of accessible states is maximum. Since the maximum of the entropy lies approximately at $E=0$ on the

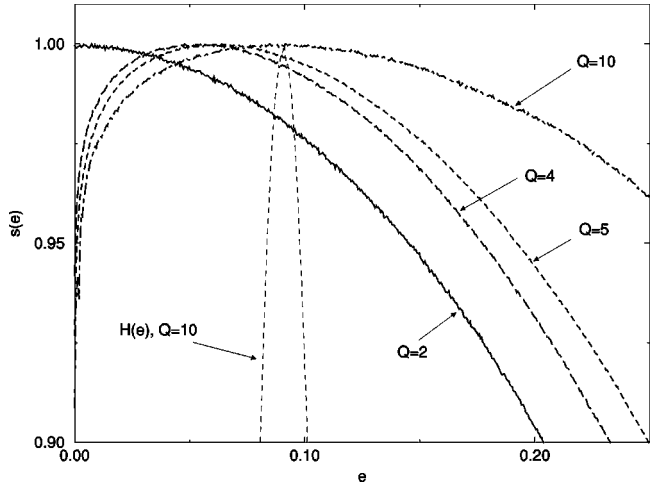

FIG. 4. Normalized entropy of the system as a function of the energy e for different numbers of partitions Q, evidencing the differences in their maximum position. We consider an instance with 120 numbers. The fundamental difference between bi- and multipartitioning is that for the latter the maximum of the entropy deviates from the origin, which turns out to be a minimum, making it more difficult to find nearly optimal solutions. The thin-dashed line is the normalized histogram of visits for a random spin-flip algorithm ($Q=10$).

NPP, any algorithm which performs random movements will find reasonable solutions for such problem.

For $Q>2$ we have a completely different scenario. The maximum of the entropy is not near $E=0$, indeed, $E=0$ is a minimum and the number of solutions with $E\approx 0$ decreases, at least, exponentially with Q. Differently from the $Q=2$ case, an algorithm based on random movements would drive the system away from the ground state. In order to illustrate this fact, we also show in Fig. 4 the normalized histogram of visits for the $Q=10$ case, where 2.4×10^8 random flips were made on the spins s. The effects of this behavior of the entropy on the construction of new algorithms should be taken into account. Based on the traditional differencing scheme [14–16] we can say that, if the number of nearly optimal solutions increases exponentially as $E\rightarrow 0$, it is always possible to find better and better solutions, the computational time spent on the search being the only barrier. For the $Q>2$ NPP this kind of procedure does not seem to work so efficiently, since the number of solutions *decreases* exponentially for $E\rightarrow 0$.

The approximate values of the entropy $S(E)=\ln g(E)$ we have found for $E\approx 0$ (considering a window of size $\epsilon = E_{MAX}/1024$) were 77.76, 129.04, 150.45, 187.03, and 484.87, for $Q=2$, 3, 4, 5, and 10, respectively. For $Q=2$ it is possible to calculate analytically the value of $S(0)$ once we know $\langle a^2\rangle$ (see Refs. [5,11]). The expected value for the particular instance considered as an example is $S(0)=77.06$.

V. CONCLUSIONS

We showed in this paper that the multicanonical method for obtaining thermodynamic averages of statistical systems can provide a tool for assessing the complexity of computer science problems, such as the number partitioning problem (NPP). This problem is one of the six basic computer science problems, according to Garey and Johnson [12], and can be formulated as a spin-glass problem. Based on this analogy we proposed a statistical mechanics method for computing the spectral degeneracy of the NPP problem which gives us information about the number of solutions for a given cost E and cardinality difference m. We have studied an extension of this problem for Q partitions and observed a fundamental difference between the classical ($Q=2$) and the generalized ($Q>2$) NPP, which explains why it is so difficult to find good solutions for the latter case. This information can be very useful in the construction of new algorithms.

ACKNOWLEDGMENTS

The authors wish to thank the International Center for Theoretical Physics (ICTP) in Trieste, Italy for financial support. This work had begun while attending the ''School on Statistical Physics Methods Applied to Computer Science'' at ICTP. We also thank F. F. Ferreira and J. F. Fontanari for useful discussions, and J.F. Stilck and T.J.P. Penna for a careful reading of the manuscript. Both authors acknowledge financial support from Brazilian agencies CNPq and CAPES.

[1] *Biologically Inspired Physics*, edited by L. Peliti, Vol. B263 of *NATO ASI Series B: Physics* (Plenum, New York, 1991).

[2] S. M. de Oliveira, P. M. C. de Oliveira, and D. Stauffer, *Evolution, Money, War and Computers* (Teubner-Text, Leipzig, 1999).

[3] P. Cheeseman , B. Kanefsky, and W. M. Taylor , in *Proceedings of IJCAI-91*, edited by J. Mylopoulos and R. Rediter (Morgan Kaufmann, San Mateo, CA, 1991).

[4] I. P. Gent and T. Walsh, in *Proceedings of the 8th International Symposium on Artificial Intelligence* (ITESM, Monterrey, 1995), pp. 356–364.

[5] S. Mertens, Phys. Rev. Lett. **81**, 4281 (1998).

[6] F. F. Ferreira and J. F. Fontanari, J. Phys. A **31**, 3417 (1998).

[7] R. Monasson, R. Zecchina S. Kirkpatrick, B. Selman, and L. Troyansky, Nature (London) **400**, 133 (1999).

[8] F. F. Ferreira and J. F. Fontanari, Physica A **289**, 54 (1999).

[9] F. F. Ferreira and J. F. Fontanari, e-print cond-mat/9910525.

[10] S. Mertens, Phys. Rev. Lett. **84**, 1347 (2000).

[11] S. Mertens, e-print cond-mat/0009230.

[12] M. R. Garey and D. S. Johnson, *Computer and Intractability: A Guide to the Theory of NP-Completeness* (Freeman, San Francisco, CA, 1979).

[13] B. A. Berg and T. Neuhaus, Phys. Lett. B **267**, 249 (1991); B. A. Berg, Int. J. Mod. Phys. C **3**, 1083 (1992).

[14] N. Karmarkar and R. M. Karp, Technical Report No. UCB/CSD 82/113, Computer Science Division, University of California, Berkeley, 1982 (unpublished).

[15] R. E. Korf, Artif. Intell. Eng. **106**, 181 (1998).

[16] S. Mertens, e-print csds/990311.

[17] M. Mézard, G. Parisi, and M. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).

[18] E. G. Coffman and G. S. Lueker, *Probabilistic Analysis of Packing and Partitioning Algorithms* (John Wiley and Sons, New York, 1991).

[19] L.-H. Tsai, SIAM J. Comput., **21**, 59 (1992).

[20] M. Y. Choi, H. Y. Lee, and S. H. Park, J. Phys. A **30**, L748 (1997).

[21] G. Besold, J. Risbo, and O. G. Mouritsen, Comput. Mater. Sci. **15**, 311 (1999).

[22] J. Lee and M.Y. Choi, Phys. Rev. E **50**, R651 (1994).

[23] B. A. Berg, in *Proceedings of the International Conference on Multiscale Phenomena and Their Simulations, Bielefeld, October 1996*, edited by F. Karsch, B. Monien, and H. Satz (World Scientific, Singapore, 1997); W. Janke, Physica A **254**, 164 (1998); B. Dünweg, in *Monte Carlo and Molecular Dynamics of Condensed Matter Physics, Como, July 1995*, edited by K. Binder and C. Ciccotti (Societá Italiana di Fisica, Bologna, 1996), p. 215; B. A. Berg, U. H. E. Hansman, and T. Heuhaus, Phys. Rev. B **47**, 497 (1993).

[24] J. Lee, Phys. Rev. Lett. **71**, 211 (1993).

[25] B. A. Berg, U. H. E. Hansmann, and Y. Okamoto, J. Phys. Chem. **99**, 2236 (1995).

[26] I. Shteto, J. Linares, and F. Varret, Phys. Rev. E **56**, 5128 (1997).

[27] M. E. J. Newman and G. T. Barkema, *Monte Carlo Methods in Statistical Physics* (Oxford, New York, 1999).