Stable propagation algorithm for the minimization of the Bethe free energy

INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS A: MATHEMATICAL AND GENERAL

J. Phys. A: Math. Gen. **36** (2003) 11201–11211 PII: S0305-4470(03)59695-0

Stable propagation algorithm for the minimization of the Bethe free energy

M Pretti and A Pelizzola

Istituto Nazionale per la Fisica della Materia (INFM) and Dipartimento di Fisica,
Politecnico di Torino, Corso Duca degli Abruzzi 24, I-10129 Torino, Italy

Received 17 February 2003, in final form 4 July 2003
Published 22 October 2003
Online at stacks.iop.org/JPhysA/36/11201

Abstract

We propose a new propagation algorithm for the minimization of the Bethe free energy for a generic lattice model with pair interactions. The algorithm turns out to be more stable than belief propagation, as it reaches a fixed point also for highly frustrated systems such as spin glasses, and faster than the provably convergent double loop algorithms.

PACS numbers: 05.10.−a, 05.50.+q, 89.70.+c

1. Introduction

The Bethe approximation [1] is a well-known technique in the statistical mechanics of lattice models. Basically, it improves the ordinary mean-field theory, by taking into account correlations between nearest-neighbour (NN) sites, but can be presented under different points of view. In the original work by Bethe [2], it has been introduced as a self-consistent field theory. Subsequently, it has been formulated [3] as a simplified (‘quasi-chemical’) evaluation of the number of configurations of the system, hence of its entropy. Moreover, it can be seen as the lowest step (pair approximation [1, 4]) of a hierarchy of approximations that take into account correlations up to arbitrarily large clusters, which is known as the cluster variation method [5, 6]. The last formulation shows most clearly the variational nature of the Bethe approximation, which determines thermodynamic equilibrium states of a system as the minima of a suitable approximate free energy (the Bethe free energy), whose variational parameters are single site and NN pair probability distributions (PDs).

Recent works [7–9] have attracted new interest in the Bethe approximation, as they have shown that the highly successful belief propagation (BP) algorithm [11], employed for solving statistical inference problems on generic graphical models [12–14], actually coincides with the minimization of a Bethe free energy. Statistical inference includes a wide range of problems of technological relevance such as image restoration [15], artificial vision [13], decoding of error-correcting codes [12], diagnosis [16]. The inference problem can be generally mapped onto a thermodynamic system defined on a graph [17], and ultimately amounts to the determination of the Boltzmann distribution of statistical mechanics [8]. It has long been known by statistical physicists [18] that the Bethe approximation is exact for tree-like graphs, that is for graphs without loops. Nevertheless, it has been shown that BP algorithms, that is the Bethe approximation, work surprisingly well also for inference problems mapped on graphs with loops [14], such as decoding of the high performance turbo codes [12]. Moreover, the BP algorithm is, on the basis of the recently proposed 'survey propagation' method [19, 20], a very powerful algorithm for combinatorial optimization problems which extends the ground-state version of belief propagation to cases with a complex energy landscape, exhibiting many local minima.

It must be remarked that, despite its success, the BP algorithm is known [9, 21, 22] to fail to converge in certain highly frustrated cases, a prototype of which is the Edwards–Anderson spin glass model [23]. Alternative algorithms exist, such as the natural iteration method (NIM), by the inventor of the cluster variation method himself [24–27] and the concave–convex procedure (CCCP) by Yuille [22], which lowers the free energy at each iteration, and hence converge to local minima. Nevertheless, the latter algorithms turn out to be much slower than BP [28]. In this paper we introduce a new propagation algorithm for the minimization of the Bethe free energy, and compare its performance against BP, on the Edwards–Anderson model [23]. The algorithm is based on a factorization of the model PD that holds for tree-like models (Bethe approximation), and on an exact expression for the conditioned PD of any cluster of sites with respect to its neighbourhood, whence it will be denoted as 'conditioned probability' (CP) algorithm. We have not been able to prove analytically the convergence of the algorithm, and its stability with respect to the initial condition, but both have been verified on a large range of parameter values for the model under consideration. In contrast, the BP algorithm fails to converge. This paper is organized as follows. In section 2 we introduce the new algorithm. In section 3 we recall the basics of the BP algorithm. In section 4 we compare the behaviours of the two algorithms on the spin glass model, and discuss the results. Section 5 is devoted to a short analysis of running times, in comparison with the alternative (NIM and CCCP) algorithms mentioned above, while in section 5 we give some concluding remarks. In the appendix we verify explicitly that the fixed points of the BP algorithm are also fixed points of the CP algorithm.

2. The CP algorithm

Let us first introduce the basic idea of the algorithm. Consider a model, defined on a graph, such that a discrete variable is associated with each site (node) of the graph. According to figure 1, let us denote by A any cluster of sites in the graph, by B the 'neighbourhood' of A (that is the set of sites interacting with the sites in A), and by C the cluster made up of all graph sites except those in A. With these assumptions, the statistical (Boltzmann) weight $w(x)$ of a configuration x of the model can be decomposed into

$$w(x) = w_{AB}(x_A, x_B) w_C(x_C) \tag{1}$$

where $w_{AB}(x_A, x_B)$ weighs only interactions among sites in A (whose configuration is denoted by x_A), and interactions between sites in A and sites in B (the configuration of the latter being denoted by x_B). Similarly $w_C(x_C)$ weighs only interactions among sites in C (configuration denoted by x_C). The PD $p_A(x_A)$ of the configuration x_A can be expressed, via the Bayes theorem, as

$$p_A(x_A) = \sum_{x_B} p_{A|B}(x_A|x_B) p_B(x_B) \tag{2}$$

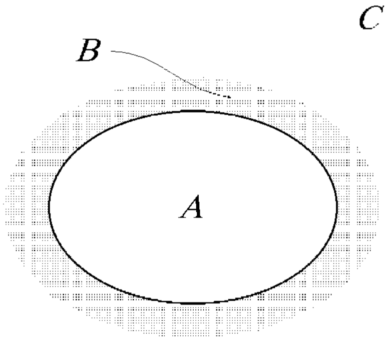

Figure 1. Sketch of the notation employed for a generic lattice: A represents any cluster of sites (a subset of the lattice); B denotes the neighbourhood of A (the set of sites interacting with A); C is the set of lattice sites that are not in A (the complement of A). Note that $A \cap B = A \cap C = \emptyset$, while $B \cap C = B$.

where $p_B(x_B)$ is the PD of x_B, and $p_{A|B}(x_A|x_B)$ is the conditioned PD of x_A over x_B, while the sum is taken over all configurations x_B. The model PD $p(x)$, after normalization, can be written as

$$p(x) = \frac{w(x)}{\sum_x w(x)} \tag{3}$$

where the sum runs over all system configurations, and it is easy to show that

$$p_{A|B}(x_A|x_B) = \frac{w_{AB}(x_A, x_B)}{\sum_{\tilde{x}_A} w_{AB}(\tilde{x}_A, x_B)}. \tag{4}$$

Replacing equation (4) into equation (2) provides a relationship between the cluster A PD and the PD of its neighbourhood $p_B(x_B)$. If A is taken to be a single site and B is the set of its NNs (provided the model includes NN interactions only), while, as an approximation, $p_B(x_B)$ is assumed to factorize into a product of single site PDs, then one obtains a self-consistent equation for single site PDs. For Ising-like models this approach has been known as 'hard spin' mean-field theory [29–32].

Let us now consider a model incorporating only NN interactions. In this case, Boltzmann weights can be written as

$$w(x) = \prod_{\langle ij \rangle} w_{ij}(x_i, x_j) \prod_i [w_i(x_i)]^{1-q_i} \tag{5}$$

where the former product runs over NN pairs, and the latter over sites. Moreover, $w_{ij}(x_i, x_j)$ weighs interactions between sites i and j and of both with external fields (if present), $w_i(x_i)$ weighs interactions of site i with external fields only, and q_i is the coordination number of site i. We assume that A is any NN pair ij, and accordingly B is the $\{i'j'\}$ neighbourhood, where, from now on, i' will always run over NNs of j except i, while j' will run over NNs of i except j. The notation is explained in figure 2. The weight factor $w_{AB}(x_A, x_B)$, now denoted by $w_{ij\{i'j'\}}(x_i, x_j, \{x_{i'}, x_{j'}\})$, can then be chosen to be

$$w_{ij\{i'j'\}}(x_i, x_j, \{x_{i'}, x_{j'}\}) = w_{ij}(x_i, x_j) \prod_{i'} \frac{w_{ji'}(x_j, x_{i'})}{w_j(x_j)} \prod_{j'} \frac{w_{ij'}(x_i, x_{j'})}{w_i(x_i)}. \tag{6}$$

According to the Bethe approximation, the PD $p(x)$ is assumed to factorize in the same way as the weight factor $w(x)$ in equation (5), that is

$$p(x) = \prod_{\langle ij \rangle} p_{ij}(x_i, x_j) \prod_i [p_i(x_i)]^{1-q_i} \tag{7}$$

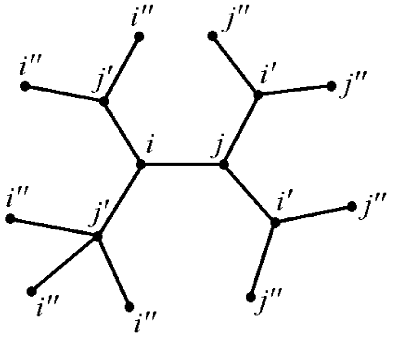

Figure 2. Sketch of the notation employed for a generic tree: i and j are a pair of NN sites; i' (j') labels NNs of j (i) except i (j); i'' (j'') labels NNs of j' (i') except i (j).

where $p_{ij}(x_i, x_j)$ denotes pair PDs, and $p_i(x_i)$ site PDs. Such a factorization turns out to be exact on a tree graph (in the Bethe approximation scheme for a given lattice the approximating tree should have the same local structure and interactions as those of the system it aims to approximate [33]). With the above factorization, the joint PD of the pair ij and its neighbourhood $\{i'j'\}$ can be written as

$$p_{ij\{i'j'\}}(x_i, x_j, \{x_{i'}, x_{j'}\}) = p_{ij}(x_i, x_j) \prod_{i'} \frac{p_{ji'}(x_j, x_{i'})}{\sum_{\tilde{x}_{i'}} p_{ji'}(x_j, \tilde{x}_{i'})} \prod_{j'} \frac{p_{ij'}(x_i, x_{j'})}{\sum_{\tilde{x}_{j'}} p_{ij'}(x_i, \tilde{x}_{j'})} \tag{8}$$

where site PDs $p_i(x_i)$ have been replaced by their expressions as marginal distributions of different pair PDs. Such expressions turn out to be essential for the algorithm to work, due to the fact that, during a run, pair PDs do not necessarily give the same marginal distribution for the same site. The neighbourhood $\{i'j'\}$ PD can be easily derived as

$$p_{\{i'j'\}}(\{x_{i'}, x_{j'}\}) = \sum_{x_i, x_j} p_{ij\{i'j'\}}(x_i, x_j, \{x_{i'}, x_{j'}\}). \tag{9}$$

As a consequence, equation (2) together with (4) is finally modified into

$$p_{ij}(x_i, x_j) = \sum_{\{x_{i'}, x_{j'}\}} \frac{w_{ij\{i'j'\}}(x_i, x_j, \{x_{i'}, x_{j'}\})}{\sum_{\tilde{x}_i, \tilde{x}_j} w_{ij\{i'j'\}}(\tilde{x}_i, \tilde{x}_j, \{x_{i'}, x_{j'}\})} p_{\{i'j'\}}(\{x_{i'}, x_{j'}\}) \tag{10}$$

which, together with equations (8) and (9), provides a set of self-consistent equations for all pair PDs in the system. The CP algorithm consists in the iterative (fixed point) solution of such set of equations.

3. The BP algorithm

Let us now briefly recall the basic idea underlying the BP algorithm. The Bethe free energy, exact for tree models, can be written as

$$\beta F = \sum_{\langle ij \rangle} \sum_{x_i, x_j} p_{ij}(x_i, x_j) \ln \frac{p_{ij}(x_i, x_j)}{w_{ij}(x_i, x_j)} + \sum_i (1 - q_i) \sum_{x_i} p_i(x_i) \ln \frac{p_i(x_i)}{w_i(x_i)} \tag{11}$$

where $\beta = 1/kT$ is the inverse temperature, while the other notation are consistent with those introduced previously. Equilibrium (pair and site) PDs are determined as those which minimize F, subject to normalization and 'compatibility' constraints. The latter impose that site PDs can be obtained as marginal distributions of pair PDs, that is,

$$p_i(x_i) = \sum_{x_j} p_{ij}(x_i, x_j) \tag{12}$$

where j denotes any NN site of i. According to the Lagrange multiplier method, one defines a suitable 'extended' free energy functional, which depends on additional unknowns (the Lagrange multipliers), but coincides with F when the constraints are satisfied. The constrained problem is usually solved in two steps, by minimizing the extended free energy with respect to PDs for given multipliers, and then by determining the latter in order to satisfy the constraints. As far as the Bethe free energy is concerned, the former step can be worked out analytically, leading to

$$\begin{aligned} \frac{p_{ij}(x_i, x_j)}{w_{ij}(x_i, x_j)} &\propto \prod_{i'} m_{i'j}(x_j) \prod_{j'} m_{j'i}(x_i) \\ \frac{p_i(x_i)}{w_i(x_i)} &\propto \prod_{j} m_{ji}(x_i) \end{aligned} \tag{13}$$

where $m_{ji}(x_i)$ are derived by a suitable mapping of the Lagrange multipliers associated with the compatibility constraints, and are known as 'messages' in the framework of the BP algorithm. The precise form of the mapping is irrelevant for our presentation. Note that proportionality symbols hide the normalization multipliers. It is possible to show that the BP algorithm corresponds to the following update rule for the messages,

$$\hat{m}_{ji}(x_i) = m_{ji}(x_i) \frac{\sum_{x_j} p_{ij}(x_i, x_j)}{p_i(x_i)} \tag{14}$$

where a hat denotes the 'new' messages, PDs are given by equations (13), while normalization is easily resolved at each step. Let us note that the above form is different from that in which the BP rule is usually reported in the literature [8, 22]. It has the advantage of showing immediately that the algorithm can just converge, if it does, to compatible PDs, proving that its fixed points are equivalent to stationary points of the Bethe free energy. By substituting equations (13) into equations (14) it is easy to rederive the usual BP rule

$$\hat{m}_{ji}(x_i) \propto \sum_{x_j} \frac{w_{ij}(x_i, x_j)}{w_i(x_i)} \prod_{i'} m_{i'j}(x_j). \tag{15}$$

4. Algorithm test on a spin glass model

In this section we give an example of the performance of the CP algorithm, compared to the BP algorithm, for the Edwards–Anderson model [23], i.e. a spin glass model with equal probability of ferro- and antiferromagnetic interactions, in zero external magnetic field. This is a highly frustrated system, for which the BP algorithm at low enough temperature does not converge, showing an apparently chaotic behaviour. We have considered a finite (10×10) square lattice of Ising spins ($s = \pm 1$), assuming NN interactions of fixed (J) intensity, with a randomly generated sign and periodic boundary conditions. Of course, this is a particular instance of the (finite) random system, for which we can approximately compute magnetization and free energy, by minimizing the Bethe functional (11). At low temperature, we find a non-uniform (glass-like) phase with finite average magnetization, and different local minima, separated by large free energy barriers. The CP algorithm turns out to converge easily to such local minima. It has to be remarked here that, as far as the square lattice model is concerned, the physical

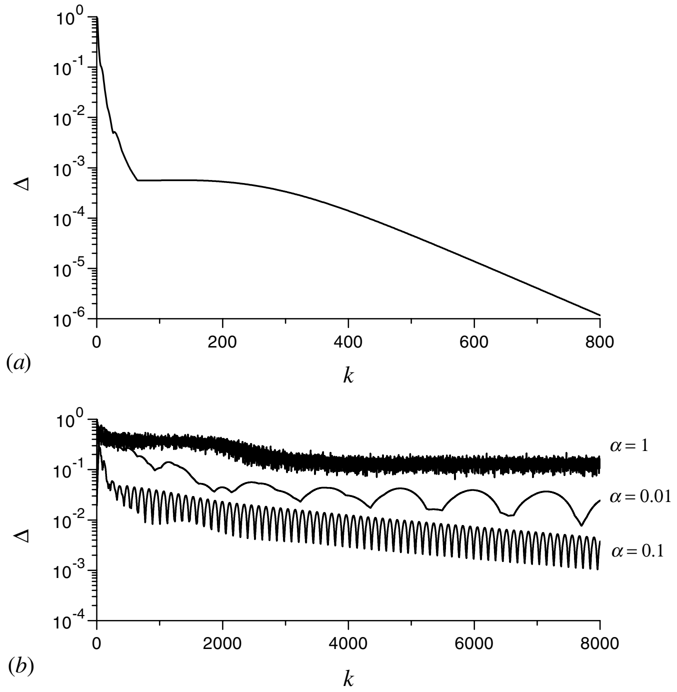

Figure 3. Δ parameter as a function of the iteration index k, for the CP algorithm (*a*) and the BP algorithm (*b*) runs. For the latter case, α denotes the over-relaxation parameter used for each run.

result is wrong, because the two-dimensional Edwards–Anderson model is generally believed to have zero transition temperature [34]. Nevertheless, our analysis is meant to give an example of Bethe free energy in which the presence of quenched randomness prevents the BP algorithm (but not the CP algorithm) from converging. Let us also note that, rigorously speaking, the Bethe approximation *usually* gives wrong answers (since it predicts symmetry breaking also for finite systems), but the resulting phase behaviour is often a good approximation for the corresponding infinite system.

In order to characterize convergence, we have to define an indicator (Δ) of the distance of the current iteration from the solution. At each iteration we estimate magnetization at site i, by means of all the possible joint PDs of i with its NNs j, in the following way:

$$M_i^{(j)} = \sum_{s=\pm 1} s \sum_{s'=\pm 1} p_{ij}(s, s'). \tag{16}$$

In principle, the estimates $M_i^{(j)}$ are different even for the same site i, because pair PDs do not satisfy the compatibility constraints during the run. We then define Δ as the maximum dispersion of magnetization estimates, that is

$$\Delta \equiv \max_i \max_j \left| M_i^{(j)} - \frac{1}{q_i} \sum_{j'} M_i^{(j')} \right| \tag{17}$$

where j and j' run over all NNs of i. In figure 3(*a*) we report Δ as a function of the iteration index k, for our model at an inverse temperature $\beta J = 0.7$, which is already in the low-temperature (glassy) regime. It is possible to see that, after the very beginning, where some

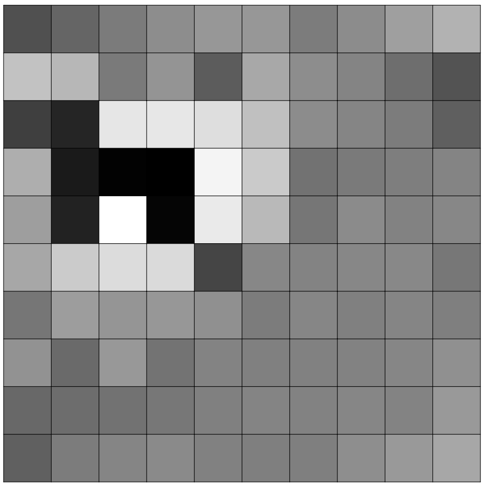

Figure 4. Grey-scale representation of site magnetization for the instance of the Edwards–Anderson model used for the CP algorithm run described in the text ($\beta J = 0.7$). Each square represents one of the 10 × 10 spins. Lighter grey tones denote higher magnetizations.

oscillations are present, Δ decreases in a regular exponential way, showing the algorithm convergence. We have observed a similar behaviour also for much larger systems and lower temperatures. For the particular run reported here, the initial condition has been chosen to be the one with statistically independent site PDs with uniform magnetization $M_i = 0.9$ for all i. Anyway, it has been possible to verify that initial conditions do not affect convergence, even if different local minima can be reached from different starting points. The local magnetizations we have obtained with this run are plotted in figure 4 in grey scale.

Let us now consider the performance of the BP algorithm on exactly the same system at the same temperature. For this case we define Δ in a slightly different way. In the BP algorithm, estimates of both pair and single site PDs are available at each iteration. So we can also evaluate the site magnetizations directly, making use of the site PDs, in the following way:

$$M_i = \sum_{s=\pm 1} s p_i(s). \tag{18}$$

It seems natural to define magnetization dispersion with respect to the latter quantity, that is

$$\Delta = \max_i \max_j \left| M_i^{(j)} - M_i \right|. \tag{19}$$

From figure 3(b) it is evident that, even for this quite small system, the BP algorithm turns out not to converge, indeed it shows an apparently chaotic behaviour. We have investigated the possibility of introducing an 'over relaxation' parameter $\alpha < 1$, that is to give an exponent α to the probability ratio in equation (14), in order to reduce oscillations and favour convergence. Nevertheless, figure 3(b) shows that, upon decreasing α, only the frequency of oscillations is monotonically reduced, but for instance the amplitude is higher for $\alpha = 0.01$ than for $\alpha = 0.1$, and a convergent behaviour is never obtained.

Let us finally note that we have not been able to prove in general the convergence of the CP algorithm. For optimization algorithms aimed at minimizing functions which are bounded from below, such a proof usually uses the fact that the cost function (free energy) is reduced at each step [22, 24, 35]. This cannot be the case for the CP algorithm, because the free energy is

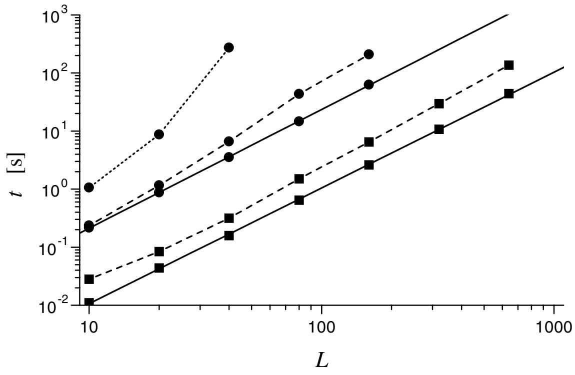

Figure 5. Running times for the BP (squares) and the CP algorithm (circles) on the homogeneous $L \times L$ Ising ferromagnet (solid lines), with 1% (dashed lines) or with 10% (dotted line) random antiferromagnetic bonds, as a function of L.

not a well-defined quantity when, during execution, pair PDs do not satisfy the compatibility constraints.

5. Execution times and scaling

Having shown that CP converges even when BP fails, it is interesting to compare the running times of CP with those of the double loop algorithms, NIM and CCCP, which are known to converge in every case. For this analysis, we have actually considered a slightly different system, that is an $L \times L$ square lattice Ising model, with randomly chosen 1% antiferromagnetic bonds, and periodic boundary conditions. Temperature is given by $\beta J = 0.5$ and initial conditions correspond to a nearly saturated ferromagnetic configuration, namely $M_i = 0.999\,999$, $\forall i$. The procedures have been terminated when $\Delta < 10^{-6}$. The reason for choosing a test model with very little disorder (so that also BP turns out to converge) is due to the fact that taking into account only single instances of the random system introduces a 'noise' in the evaluation of execution times. In the case of highly disordered systems, such a noise does not allow us to recognize the actual scaling behaviour of the algorithms, with respect to the system dimension. This effect can be clearly observed in figure 5, where we compare execution times for systems with different degrees of disorder. Only the limit of a homogeneous ferromagnetic system displays a precise $t \propto L^2$ scaling. On an ordinary DEC Alpha workstation, for CP we obtain $t/L^2 \approx 1.9 \times 10^{-3}$s, while BP gives $t/L^2 \approx 1.1 \times 10^{-4}$s, that is more than one order of magnitude faster. It is interesting to observe how a little disorder can increase running times. Namely, in the 1% case we obtain a factor ranging from 1 to 3 for CP, and from 2.5 to 3 for BP. Having performed the runs on exactly the same systems with the same initial conditions, this fact confirms the different nature of the two algorithms.

Figure 6 compares the running times of BP, CP, NIM and CCCP. In spite of the presence of disorder, times generally keep on scaling roughly as L^2, but this is not true for the NIM, which scales roughly as $L^{3.6}$. As previously mentioned, BP is the fastest algorithm, but CP is roughly an order of magnitude faster than CCCP, which in turn is faster than NIM (as already shown in a one-dimensional case [28]). Therefore, it turns out that CP is the fastest one among the always converging algorithms.

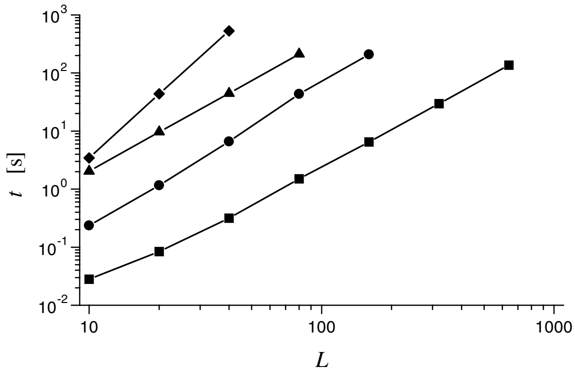

Figure 6. Running times for the BP (squares), CP (circles), CCCP (triangles), and NIM algorithm (diamonds), on the $L \times L$ Ising ferromagnet with 1% random antiferromagnetic bonds, as a function of L.

6. Conclusions

In this paper we have proposed a new (CP) algorithm for the minimization of the Bethe free energy. The CP algorithm is based on two main ideas, that are an exact expression for the conditioned probability (CP) of a pair of sites with respect to its neighbourhood, and the factorization of the total system PD in terms of pair and site PDs. Such a factorization, exact for tree graphs, just gives rise to the Bethe expression for the free energy. This guarantees that already in principle the CP algorithm (if convergent) finds precisely the stationary points of the Bethe free energy. This is also shown analytically in the appendix, while the fact that it converges to the minima can be (by now) only verified numerically. A very appealing feature of the CP algorithm is that it seems to converge quite easily also for highly frustrated spin glass models, for which the previously known BP algorithms fail. Such a property might be of use especially in the framework of the survey propagation method [19, 20], dealing with disordered systems. We have shown this feature, which—we remark—we have not been able to prove in general, on a simple instance of the Edwards–Anderson model at a fixed temperature, but we have verified it on many other cases. With respect to another recently proposed algorithm, the CCCP, which has been proved analytically to converge to the minima of the Bethe free energy, as well as the NIM, the CP algorithm has the advantage of being a single loop algorithm, while the others are double loop ones. This makes the CP much faster than CCCP and NIM, as we have shown. Let us finally remark that it would be of interest to extend the present work to higher order approximations of the cluster variation method, and/or to models with interactions not limited to NNs, and work is in progress along these lines.

Appendix. Equivalence of CP and BP fixed points

The equivalence between the fixed points of the CP algorithm and the minima of the Bethe free energy should be guaranteed by the thermodynamic variational principle, and by the fact that a completely general free energy

$$\beta F = \sum_{x} p(x) \log \frac{p(x)}{w(x)} \tag{A1}$$

turns into the Bethe free energy (11) when factorizations (5) and (7) hold. Nevertheless it may be interesting to verify this issue explicitly. Indeed, we show that inserting the BP rules (equations (13) and (14)) into the self-consistent equation which defines our method (equation (10)) gives rise to an identity.

According to equations (8) and (9), when the compatibility constraints (12) are verified (which is true at a fixed point of the BP algorithm) we can write

$$p_{\{i'j'\}}(\{x_{i'}, x_{j'}\}) = \sum_{x_i, x_j} p_{ij}(x_i, x_j) \prod_{i'} \frac{p_{ji'}(x_j, x_{i'})}{p_j(x_j)} \prod_{j'} \frac{p_{ij'}(x_i, x_{j'})}{p_i(x_i)}. \tag{A2}$$

By replacing the PDs with their expressions as a function of the messages, equations (13), and taking into account equation (6), we find

$$\frac{p_{\{i'j'\}}(\{x_{i'}, x_{j'}\})}{\sum_{x_i, x_j} w_{ij\{i'j'\}}(x_i, x_j, \{x_{i'}, x_{j'}\})} \propto \prod_{i'} \prod_{j''} m_{j''i'}(x_{i'}) \prod_{j'} \prod_{i''} m_{i''j'}(x_{j'}) \tag{A3}$$

where the inner products run respectively over all sites j'' that are NNs of i' except j, and over all sites i'' that are NNs of j' except i (see figure 2). Let us now substitute the latter expression into the self-consistent equation (10). By means of some simple algebra we obtain

$$\frac{p_{ij}(x_i, x_j)}{w_{ij}(x_i, x_j)} \propto \prod_{i'} \sum_{x_{i'}} \frac{w_{ji'}(x_j, x_{i'})}{w_j(x_j)} \prod_{j''} m_{j''i'}(x_{i'}) \prod_{j'} \sum_{x_{j'}} \frac{w_{ij'}(x_i, x_{j'})}{w_i(x_i)} \prod_{i''} m_{i''j'}(x_{j'}). \tag{A4}$$

Finally, from equation (15) we see that, at a fixed point of the BP algorithm, we have

$$m_{ji}(x_i) \propto \sum_{x_j} \frac{w_{ij}(x_i, x_j)}{w_i(x_i)} \prod_{i'} m_{i'j}(x_j) \tag{A5}$$

which, replaced into the previous equation, give rise again to equation (13), thus to an identity.

References

[1] Burley D M 1972 *Phase Transitions and Critical Phenomena* vol 2 ed C Domb and M S Green (New York: Academic) ch 9

[2] Bethe H A 1935 *Proc. R. Soc.* A **150** 552

[3] Guggenheim E A 1952 *Mixtures* (Oxford: Oxford University Press)

[4] Kikuchi R 1951 *Phys. Rev.* **81** 988

[5] An G 1988 *J. Stat. Phys.* **52** 727

[6] 1994 Foundations and applications of cluster variation method and path probability method *Prog. Theor. Phys. Suppl.* **115**

[7] Kabashima Y and Saad D 1998 *Europhys. Lett.* **44** 668

[8] Yedidia J S 2001 *Advanced Mean Field Methods: Theory and Practice* ed M Opper and D Saad (Cambridge, MA: MIT Press)

[9] Yedidia J S, Freeman W T and Weiss Y 2001 *Advances in Neural Information Processing Systems 13* ed T K Leen, T G Dietterich and V Tresp (Cambridge, MA: MIT Press)

[10] Kabashima Y 2002 *Preprint* cond-mat/0211500

[11] Pearl J 1988 *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (San Mateo, CA: Morgan Kaufmann)

[12] McEliece R J, MacKay D J C and Cheng J F 1998 *IEEE J. Sel. Areas Commun.* **16** 140

[13] Freeman W T, Pasztor E C and Carmichael O T 2000 *Int. J. Comput. Vis.* **40** 25

[14] Murphy K P, Weiss Y and Jordan M I 1999 *Proc. 15th Conf. on Uncertainty in AI* (San Mateo, CA: Morgan Kaufmann)

[15] Tanaka K 2002 *J. Phys. A: Math. Gen.* **35** R81

[16] Kappen H J 2002 *Modeling Bio-medical Signals* (Singapore: World Scientific)

[17] Smyth P 1997 *Pattern Recogni. Lett.* **18** 1261

[18] Baxter R J 1982 *Exactly Solved Models in Statistical Mechanics* (New York: Academic)

[19] Mézard M, Parisi G and Zecchina R 2002 *Science* **297** 812
[20] Mézard M and Zecchina R 2002 *Phys. Rev.* E **66** 056126
[21] Kappen H J and Wiegerinck W 2002 *Advances in Neural Information Processing Systems* vol 14 ed T G Dieterich, S Becker and Z Ghahramani (Cambridge, MA: MIT Press)
[22] Yuille A L 2002 *Neural. Comput.* **14** 1691
[23] Edwards S F and Anderson P W 1975 *J. Phys. F: Met. Phys.* **5** 965
[24] Kikuchi R 1974 *J. Chem. Phys.* **60** 1071
[25] Kikuchi R 1976 *J. Chem. Phys.* **65** 4545
[26] Kikuchi R, Kokubun H and Katsura S 1986 *J. Phys. Soc. Japan.* **55** 1836
[27] Pelizzola A 1994 *Physica* A **211** 107
[28] Pelizzola A *Phys. Rev.* E submitted
[29] Netz R R and Berker A N 1991 *Phys. Rev. Lett.* **66** 377
[30] Banavar J R, Cieplak M and Maritan A 1991 *Phys. Rev. Lett.* **67** 1807
[31] Pelizzola A and Pretti M 1999 *Phys. Rev.* B **60** 10134
[32] Kabakçioḡlu A 2000 *Phys. Rev.* E **61** 3366
[33] Gujrati P D 1995 *Phys. Rev. Lett.* **74** 809
[34] Houdayer J 2001 *Eur. Phys. J.* B **22** 479
[35] Pretti M 2003 *J. Stat. Phys.* **111** 993