Loopy belief propagation: Benefits and pitfalls on Ising-like systems

N. Stoop, T. Ott, and R. Stoop

Institute of Neuroinformatics ETHZ/UZH
Winterthurerstr. 190, 8057 Zürich, Switzerland
Email: norbert@ini.phys.ethz.ch, tott@ini.phys.ethz.ch, ruedi@ini.phys.ethz.ch

Abstract—Belief propagation (BeP) is a candidate for the accelerated evaluation of statistical averages if compared to Monte-Carlo approaches. For binary systems on infinite grids or with periodic boundary conditions (regular grids), we investigate the physical fixed points and their stability. Critical slowing down of the method is observed at the 2nd-order phase transition with $T_C \approx 2.89$. Above T_C, convergence is guaranteed. Below the critical temperature T_C, BeP convergence depends dramatically on the choice of initial conditions. This leads to convergence patterns typical for fractal basin boundaries.

1. Introduction

Belief propagation (BeP) is a relatively new and powerfull method for inference and optimization problems. Inference problems arise in many fields of statistical physics, error-correcting codes and machine learning. Interestingly, some of the previously developed methods such as turbocodes or the transfer matrix approach in physics are in fact just variations of the same belief propagation technique. Inference problems deal typically with questions such as: Given a set of variables with statistical dependencies, what are their most probable states when only the states of a possibly small group of variables is known from data? As an example, think of the following system, taken from [1]: Holidays in an Asian country (A) increase the risk of tuberculosis (T), whereas smoking (S) can cause lung cancer (L) or bronchitis (B). Either (E) tuberculosis or lung cancer can be detected by an X-ray analysis (X), which, however, can not distinguish between both illnesses. Shortness of breath (Dyspnoea, D) can either be caused by bronchitis (B) or either (E) lung cancer or tuberculosis (see Fig 1).

Such systems are called Bayesian networks and can be represented by graphs, where each variable corresponds to a node and the dependency between variables is denoted by interconnecting lines (”edges”). In our example, the dependencies between nodes are *directed*. Smoking increases the probability of lung cancer, but not vice-versa. In the context of Fig. 1, two probabilities are of main interest: The overall joint probability, $p(x)$ that any event, i.e. illness, diagnosis etc. occurs, defined as $p(x) = p(x_A)p(x_S)p(x_T|x_A)p(x_L|x_S)p(x_B|x_S) \cdot p(x_E|x_L, x_T)p(x_D|x_B, x_E)p(x_X|x_E)$. More generally, the joint probability can be written in the form $p(x) = \prod_{i=1}^{N} p(x_i|Par(x_i))$, where $Par(x_i)$ denotes the parents of node i. In many cases, however, the probability of a certain diagnosis, x_D for example, is of greater relevance. This question deals with the marginal probabilities, which are calculated by summing over all possible states of a node's parents:

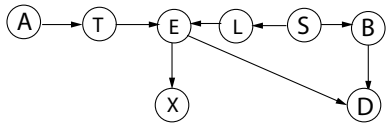

Figure 1: The directed graph example (see text).

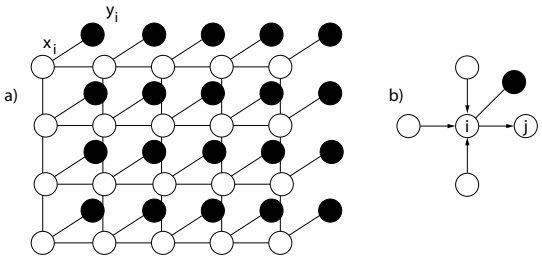

Figure 2: a) MRF graph. Black circles denote observed nodes, whereas white nodes are unknown. b) Message update rule for the belief propagation algorithm.

$$p(x_N) = \sum_{x_1} \sum_{x_2} \dots \sum_{x_{N-1}} p(x_1, x_2, \dots x_N) \qquad (1)$$

From this definition, it is obvious that an exact mathematical computation is only practical for small graphs, since the number of terms in the sum grows exponentially with the number of nodes.

Pairwise Markov Random Fields (MRF) and Ising models: Consider the computer graphics task to infer some quantities of the underlying scene x_i (Fig. 2 a) from pixel image data y_i. We further assume that there is a statistical relation $\phi(x_i, y_i)$ between x_i and y_i for each pixel i. This function is often called *evidence* for x_i, as it influences the probability for a scene site i to be of value x_i, given that one

observes y_i. Additionally, in order to deduce anything at all from the image, there has to exist an underlying structure in the scene, expressed by a *compatibility function* $\psi_{ij}(x_i, x_j)$. The compatibility function can be understood as a coupling between the scene quantities x_i (the word "coupling" is used here as it turns out that ψ_{ij} reduces to the physical coupling constant J for Ising systems).

Given these quantities and functions, we can write the overall joint-probability as

$$p(\{x\},\{y\}) = \frac{1}{Z}\prod_{ij}\psi_{ij}(x_i,x_j)\prod_i \phi(x_i,y_i) \qquad (2)$$

where Z is the normalization constant and the product runs over all connected x_i's (for example, the nearest neighbours on a regular grid). In contrast to the Bayesian network discussed above, the MRF is undirected and pairwise, since the compatibility functions ψ_{ij} only depend on pairs of nodes i and j. As with Bayesian networks, the computation of exact marginal probabilities is only possible for very small systems. The instructive part of MRF is how they relate to magnetic spin models. The energy of such a system is given by its spin-part Hamiltonian, $E = -\sum_{(i,j)} J_{ij} s_i \cdot s_j - \sum_i h_i s_i$, where the first sum is taken over a site's nearest neighbours and s_i is the spin of i-th site. Here, J_{ij}/T is identified essentially with $\log(\psi_{ij})$. Its physical interpretation is that of a coupling strength between sites i and j. $\log(\phi_i)$ is essentially identified with h_i/T, following the physical interpretation of an external magnetic field. From statistical physics it is known that the states of the system in the canonical ensemble obey Boltzmann's law: The probability of finding the system in some spin configuration $\{s\}$ is $p(\{s\}) = \frac{1}{Z} e^{E(\{s\})/k_B T}$. Comparing this to Eq. (2), we realize that the MRF corresponds to a spin model at temperature T. It also follows that the normalization constant Z in (2) can be interpreted as the system's partition function.

Belief propagation approach: Because of their close relation to physics, we will focus in the following on MRF. Again, we are interested in calculating joint and marginal probabilities, as given by Eq. (2), but in an approximate manner only. In the following, we will identify $\phi(x_i, y_i) \equiv \phi(x_i)$, since, for simplicity, we assume the observed nodes y_i to be fixed. The idea behind belief propagation is to approximate marginal probabilities for a site i by so-called beliefs $b_i(x_i)$. The beliefs are calculated from *messages* m_{ji} sent to the i-th site from its neighbours j:

$$b_i(x_i) = k\phi_i(x_i)\prod_{j\in N(i)} m_{ji}(x_i) \qquad (3)$$

That is, we take the product over all messages coming in from the neighbouring sites $N(i)$, multiply by the local evidence ϕ_i and normalize by a constant k (all beliefs at a site have to sum to one). The messages m_{ji} are determined self-consistently by the iterative update rule

$$m_{ij} \leftarrow \sum_{x_i} \phi_i(x_i)\psi_{ij}(x_i,x_j)\prod_{k\in N(i)/j} m_{ki}(x_i) \ . \qquad (4)$$

In other words, the new message site i is going to send to site j is determined by the messages that were previously sent to the i-th site from all its neighbours except j, weighted by the "coupling strength" ψ_{ij} and the local evidence ϕ_i (see Fig. 2 b). It is easy to prove that the beliefs $b_i(x_i)$ converge to the exact marginal probabilities $p_i(x_i)$ for singly connected graphs [1]. However, in contrast to the direct calculation of the marginals (1), computing time for belief propagation on singly connected graphs only grows linear with the number of connections between the nodes.

For loopy graphs, the situation is worse. On such graphs it is not guaranteed that belief propagation converges at all. For spin systems, it was shown that there exists a critical temperature T_C above which loopy belief propagation converges [2]. In this regime, the choice of initial messages is irrelevant and has only small influence on the required computing time. At lower temperatures, near and in the ferromagnetic phase, convergence, however, may not be achieved. It is important to note that if BeP converges, the beliefs correspond to stationary points of the Bethe free energy. This correspondence finally justifies the use of BeP for simulating thermodynamical systems such as the Ising model. A proof of this theorem can be found in [1].

2. BeP convergence and critical slowing down

The behaviour of BeP in dependence on the temperature is of great interest, e.g. for sequential superparamagnetic clustering (SSC, [3]). SSC is based on magnetic spin models and requires sweeps over large temperature ranges to come up with natural data clusters. The task can be accomplished by Monte Carlo simulations ([3]), but BeP is much faster (up to a factor of 20) - if it converges [4]. When clusters break apart, BeP slowing down is observed, where it can be shown that this temperature is exactly the critical temperature of the phase transition. To investigate this phenomenon, we will restrict ourselves to Ising systems, for they allow a message parametrization [2],

$$\tanh \nu_{ij} := m_{ij}(+1) - m_{ij}(-1) \ , \qquad (5)$$

which simplifies the update rules significantly:

$$\tanh \nu'_{ij} = \tanh J_{ij} \tanh(h_j + \sum_{k\in N(i)/j} \nu_{kj}). \qquad (6)$$

Assuming uniform coupling $J_{ij} = J$ and vanishing external field, two uniform fixed point solutions exist for Eq. (6), namely $\nu_{ij} = \nu = 0$ (zero magnetization, paramagnetic phase) and $\nu_{ij} = \nu \neq 0$ (non-vanishing magnetization, ferromagnetic phase), cf. Fig (3 a). For stability analysis, the Jacobi matrix can be calculated to be

$$\frac{\partial \nu'_{ji}}{\partial \nu_{kl}} = \frac{1-\tanh^2(h_j + \sum_{t\in N(j)/i}\nu_{tj})}{1-\tanh^2(\nu'_{ji})} \cdot \tanh J_{ij}\delta_{jl}\mathbf{1}_{N(j)/i}(k) \qquad (7)$$

It is straightforward to guess that a uniform initialization of the messages ν_{ij} is an eigenstate of the *regular* grid with periodic boundary conditions: Let us assume that each site has q nearest neighbours. We can then write the tensor (7) as a $q \cdot N \times q \cdot N$ matrix M, where N is the total number of sites, by numerating the pairs ij as rows and the pairs (kl) as columns of M. From (7) it is clear that each matrix row only contains $q-1$ non-zero entries, since every node receives $q-1$ messages only. With uniform message initializations, all $q-1$ non-zero matrix entries per row take the value

$$m := \frac{1-\tanh^2((q-1)\nu)}{1-\tanh^2(J)\tanh^2((q-1)\nu)}\tanh(J). \qquad (8)$$

Hence, the vector $(1,1,\ldots,1)$ is an eigenvector with eigenvalue $(q-1)m$. In fact, this eigenvalue is the largest one. *Proof:* Recall that a special form of *Frobenius' theorem* states that if $A_{n\times n} \geq 0$ is irreducible, each of the following is true: 1) $r = \rho(A) \in \sigma(A)$, where ρ is the spectral radius and $\sigma(A)$ the spectrum of A. 2) The unique vector p defined by $Ap = rp$, $p_i > 0$ and $\| p \|_1 = 1$ is called the *Perron* vector. 3) There are no other positive eigenvectors. Since $m > 0$, the Jacobian (7) is clearly positive semi-definite. Irreducibility of a matrix, on the other hand, is equivalent to an underlying graph being strongly connected. A strongly connected graph is a graph where each pair of nodes can be connected by a sequence of directed edges. This is obviously the case in the Ising example, since the nodes and edges form a regular grid structure. Thus, the eigenvector $(1,1,\ldots,1)$ is the Perron vector and its eigenvalue is the largest one. This result remains valid for all grids having the same number of neighbours to each node, i.e. an infinite regular grid or a finite grid with periodic boundary conditions.

Paramagnetic phase: From the above theorem it follows immediately that the fixed point $\nu = 0$ is stable as long as $J < tanh^{-1}(1/(q-1))$ (Fig (3 b)). The critical coupling/temperature is therefore $J \approx 0.347$ or $T = 1/J \approx 2.89$. This is in accordance to numerical simulations, where a typical critical slowing down can be observed in the BeP computing time (Fig. 3 c).

Moreover, it can be shown that BeP converges irrespective of the choice of initial conditions if $|\sigma(A)| < 1$ [2], where $\sigma(A)$ is the spectral radius of the matrix

$$A_{ij,kl} = \tanh(J)\delta_{il}\mathbf{1}_{N(i)/j}(k) \ , \qquad (9)$$

which is the message independent part of the Jacobi matrix (7). The condition guarantees BeP convergence for all (i.e. also non-uniform) initial conditions in the paramagnetic phase.

Ferromagnetic phase: In the ferromagnetic phase, despite having a stable, attractive fixed point (Fig. 3 b), convergence depends heavily on the choice of the initial messages ν_{ij}. Although physical solutions should break symmetry and switch to a uniform magnetization density ± 1, BeP does not necessarily converge, according to (9).

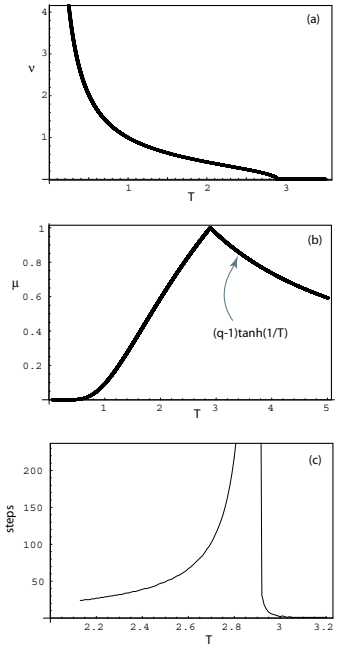

Figure 3: a) Uniform fixed points ν corresponding to the ferromagnetic and paramagnetic phases. b) Stability of the fixed points: $T < T_C (\approx 2.89)$: The ferromagnetic fixed point is stable (but not globally attractive, see text); $T > T_C$: the paramagnetic fixed point becomes stable and globally attractive. c) A typical critical slowing down is observed at the 2nd order phase transition.

One of possibly many initial conditions for which BeP does not converge in the ferromagnetic phase is a situation as shown in Fig. 4 a). In the simulations, such a vortex was considered as a perturbation of the fixed points (6) on a 3×3 regular grid with periodic boundary conditions.

In our simulations, the system was simulated for temperature sweeps below T_C. The messages were initialized to either the stable ferromagnetic fixed point (without loss of generality, $\nu > 0$ in the following) or to the (in this temperature regime) unstable $\nu = 0$ fixed point. Before the simulation was run, a perturbation of variable strength directed towards the vortex was once applied (Fig. 4 b). BeP was then started, and the iterations required for convergence were counted. If more than 500 iterations were required, the system was assumed not to converge.

The criterion for convergence was that messages would differ by less than $\epsilon = 10^{-16}$ between successive updates. The same simulation procedure was then repeated for different temperatures and perturbation strengths λ. Since the temperature range considered belongs to the ferromagnetic phase, one would expect the perturbed system to converge to the $\nu > 0$ fixed point (Fig. 4 b).

This is indeed true when perturbations are applied to the ferromagnetic fixed point. A typical convergence plot for this situation is shown in Fig. 5 a), from which we see

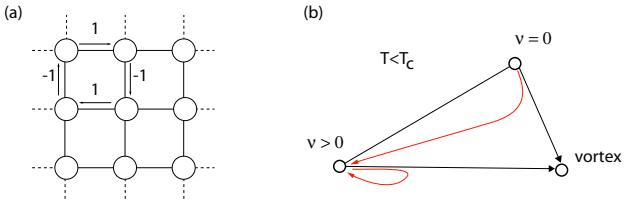

Figure 4: a) Vortex message initialization on a regular Ising grid. b) Illustration of the perturbations applied to the two fixed points. The red arrows indicate the expected behavior.

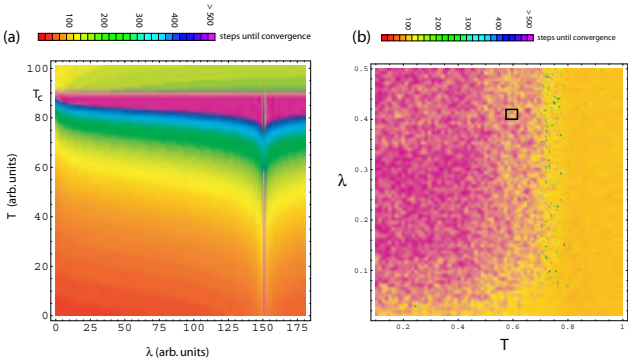

Figure 5: a) Vortex perturbation of the stable ferromagnetic $\nu > 0$ fixed point as a function of perturbation strength λ and temperature T. $\lambda = 150$ (arbitrary units) corresponds to a perturbation leading exactly to a vortex setup. b) Vortex perturbation of the unstable $\nu = 0$ paramagnetic fixed point, as a function of the perturbation strength λ and temperature T.

that BeP converges for all perturbations except for the exact vortex initialization. In the latter case, the convergence dependence as a function of temperature seems to be inverted to the general case: In the ferromagnetic phase, the vortex initialization never converges. Upon temperature increase, the vortex, however, arrives at a critical point $T'_C < T_C$, above which it always converges. Note that the vortex initialization also converges quickly at temperatures at which critical slowing down is observed in the general case. The previously mentioned "cut-off" at 500 iteration steps becomes reasonable when looking at the results in Fig. 5 a): First, a higher cut-off would only slightly narrow the peak around T_C. Second, the vortex initialization does not converge below T'_C, even if we use significantly longer simulation runs.

A completely different behaviour is observed when we start with a perturbation of the unstable paramagnetic fixed point $\nu = 0$, as shown in Fig. 5 b). Another critical temperature seems to separate convergent from non-convergent regions at $T''_C \approx 0.75$. Above this temperature, BeP converges, independently of the perturbation strength λ. Below T''_C, convergent and non-convergent initializations can be found. The behaviour is non-trivial and depends strongly on the choice of the parameters T and λ. Furthermore, as one would expect, for very small λ, BeP tends to converge more often. Apart from that, the choice of λ does not influence the behaviour on a larger scale, in contrast to the temperature.

Below T''_C, when zooming in the black square of Fig. 5 b), a self-similar structure is revealed. Even for high magnifications, a distinct boundary separating converging from non-converging regions is missing. It is an open question whether the nature of these findings lie in the theory of fractal basin boundaries.

3. Conclusion

Although loopy belief propagation on Ising systems has stable, attractive fixed points in both magnetic phases, it does generally only converge in the paramagnetic case. A BeP critical slowing down at $T_C \approx 2.89$ is observed at the phase transition, where both BeP fixed point solutions become unstable. In the ferromagnetic phase, vortex perturbations to the BeP fixed points reveal two entirely different behaviours: Perturbations of the stable ferromagnetic solution does not heavily influence BeP efficiency and convergence. With exact vortex initialization, BeP does not converge at low temperatures, whereas, interestingly, the vortex starts to converge to the paramagnetic phase at temperatures well below T_C. Perturbations of the unstable paramagnetic solution reveal a more complex situation. Convergence is usually achieved for $T > T''_C \approx 0.75$ only. At lower temperatures, self-similar patterns are observed, the nature of which yet remains to be analyzed. An appealing explanation are fractal basin boundaries, separating the convergent from the non-convergent initial conditions.

References

[1] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Understanding Belief Propagation and its generalizations", Mitsubishi Electric Research Laboratories report (2002).

[2] J. Mooij and B. Kappen, "Validity estimates for loopy Belief Propagation on binary real-world networks", NIPS 2004 Conference Proceedings (2004).

[3] T. Ott, A. Kern, W.-H. Steeb, and R. Stoop, "Sequential superparamagnetic clustering: Tracking down the most natural clusters", J. Stat. Mech., P11014 (2004).

[4] T. Ott, J. Dauwels, and R. Stoop, "Sequential superparamagnetic clustering by loopy belief propagation", Proceedings of ECCTD 2005, P328 (2005).

[5] T. Heskes, "On the uniqueness of Loopy Belief Propagation fixed points", Neural Computation **16**, 2379 (2004).