

MITSUBISHI ELECTRIC RESEARCH LABORATORIES
<http://www.merl.com>

A conversation about the Bethe free energy and sum-product (discussion document)

David J.C. MacKay, Jonathan S. Yedidia, William T. Freeman, Yair Weiss

TR2001-18 May 2001

Abstract

This discussion document records an email conversation in preparation for the Trieste meeting.

This work may not be copied or reproduced in whole or in part for any commercial purpose. Permission to copy in whole or in part without payment of fee is granted for nonprofit educational and research purposes provided that all such whole or partial copies include the following: a notice that such copying is by permission of Mitsubishi Electric Research Laboratories, Inc.; an acknowledgment of the authors and individual contributions to the work; and all applicable portions of the copyright notice. Copying, reproduction, or republishing for any other purpose shall require a license with payment of fee to Mitsubishi Electric Research Laboratories, Inc. All rights reserved.

Copyright © Mitsubishi Electric Research Laboratories, Inc., 2001
201 Broadway, Cambridge, Massachusetts 02139

Final version April 9, 2001. Made into a MERL TR May 15, 2001.

A Conversation about the Bethe Free Energy and Sum-Product

David J. C. MacKay

Department of Physics, University of Cambridge
Cavendish Laboratory, Madingley Road,
Cambridge, CB3 0HE, United Kingdom.
mackay@mrao.cam.ac.uk

Jonathan S. Yedidia & William T. Freeman

MERL,
201 Broadway
Cambridge, MA 02139, USA
yedidia/freeman@merl.com

Yair Weiss

Computer Science Division
UC Berkeley, 485 Soda Hall,
Berkeley, CA 94720-1776, USA
yweiss@cs.berkeley.edu

April 9, 2001 — Version 3.14 (final)

Abstract

This discussion document records an email conversation in preparation for the Trieste meeting.

BACKGROUND

The result that ‘belief propagation fixed-points are zero gradient points of the Bethe free energy’ (Yedidia, 2000; Yedidia *et al.*, 2000c) is an interesting one, since the sum-product algorithm (or belief propagation) has proved so useful for decoding sparse graph codes. The sort of interesting ideas motivated by this idea include:

- One could make a modified algorithm that not only has the same fixed points but provably descends the Bethe free energy (Welling and Teh, 2001; Yuille, 2001). (Sum-product does not descend a Lyapunov function; it can show chaotic behaviour.)
- Yedidia *et al.* (2000c) suggest that one could derive generalized belief propagation algorithms from better objective functions than the Bethe free energy.

But in this conversation, DJCM says “slow down, there’s a few basic points I don’t understand”.

Here are my questions, and the replies from JSY, WTF and YW

1. **Unrealisability.** In the case of the pairwise Markov network, the Bethe free energy is a function of the joint pseudo-distributions b_{ij} and the marginal pseudo-distributions b_i ,

$$F(\{b_{ij}, b_i\}) = \sum_{ij} \sum_{x_i, x_j} b_{ij}(x_i, x_j) [\ln b_{ij}(x_i, x_j) - \ln \psi_{ij}(x_i, x_j)] \\ - \sum_i (q_i - 1) \sum_{x_i} b_i(x_i) [\ln b_i(x_i) - \ln \psi_i(x_i)], \quad (1)$$

where q_i is the number of neighbours of node i . If the potentials ψ are positive then this function is well-defined for any positive setting of the pseudo-distributions $\{b\}$. The idea is that when the pseudo-distributions $\{b\}$ minimize the Bethe free energy, they will implicitly define an approximating distribution close (in some sense) to the true distribution $P(x_1, x_2, \dots, x_N)$.

However, there is a difficulty with this last step, ‘implicitly define an approximating distribution’: it is possible to set the pseudo-distributions $\{b\}$ to values such that there is *no* distribution that has those pseudo-distributions as its marginals.

Here is a simple example. Let x_1, x_2, x_3 be binary variables. Let the marginal pseudo-distributions be uniform: $b_i(x_i) = (0.5, 0.5)$; and let the joint pseudo-distributions be

$$b_{12}(x_1, x_2) = \begin{bmatrix} 0.4 & 0.1 \\ 0.1 & 0.4 \end{bmatrix} \quad (2)$$

$$b_{23}(x_2, x_3) = \begin{bmatrix} 0.4 & 0.1 \\ 0.1 & 0.4 \end{bmatrix} \quad (3)$$

$$b_{13}(x_1, x_3) = \begin{bmatrix} 0.1 & 0.4 \\ 0.4 & 0.1 \end{bmatrix} \quad (4)$$

We now prove that there can be no distribution $\{p_{000}, p_{001}, p_{010}, p_{011}, p_{100}, p_{101}, p_{110}, p_{111}\}$ that has these pseudo-distributions as its marginals. First, from the top right entry in (2), $p_{010} + p_{011} = 0.1$, so $p_{010} \leq 0.1$ and $p_{011} \leq 0.1$. Similarly, from the lower left entry in (2), $p_{100} \leq 0.1$ and $p_{101} \leq 0.1$.

Then from (3), $p_{001} \leq 0.1$ and $p_{101} \leq 0.1$ (which we already knew); and $p_{010} \leq 0.1$ (which we already knew) and $p_{110} \leq 0.1$.

Finally from (4), $p_{000} \leq 0.1$ and $p_{111} \leq 0.1$.

Thus all the joint probabilities must be less than or equal to 0.1, and the total probability (which should be one) must be less than or equal to 0.8. QED.

The fact that the arguments $\{b\}$ of the Bethe free energy can correspond to an unrealisable distribution seems a cause for concern. Can it be shown that, after minimization, the pseudo-distributions *will always* in fact be realisable? If not, what are we to make of the possibility that the optimum $\{b\}$ is unrealisable? How does this question relate to the sum-product algorithm?

[Another question one could ask is, if the pseudo-probabilities $\{b\}$ *are* realisable, how would Bethe have us define an approximate distribution that realises them? Would one, for example, imagine selecting the maximum entropy distribution that matches $\{b\}$?

JSY writes...

Yes, it is true that the b_i, b_{ij} are generally unrealizable unless the graph is a tree.

In fact, one can construct a network for which your example is the minimum of the Bethe free energy. Take

$$\psi_{12} = \begin{bmatrix} .4 & .1 \\ .1 & .4 \end{bmatrix} \quad (5)$$

$$\psi_{23} = \begin{bmatrix} .4 & .1 \\ .1 & .4 \end{bmatrix} \quad (6)$$

$$\psi_{13} = \begin{bmatrix} .1 & .4 \\ .4 & .1 \end{bmatrix} \quad (7)$$

$$\psi_1 = \psi_2 = \psi_3 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad (8)$$

in which case iterating belief propagation will give you that all the messages are $[1; 1]$, and the one and two-node beliefs are the unrealizable ones that you

JSY writes...

It is true that the Bethe free energy would be infinite if you had any terms where $b_{ij}(x_i, x_j) > 0$ when $\psi_{ij}(x_i, x_j)$ was exactly equal to zero. Fortunately, the sum-product algorithm insures that this ($b_{ij} > 0$ when it shouldn't) never happens. Since it is reasonable to take $b_{ij}(x_i, x_j) \ln \psi_{ij}(x_i, x_j) = 0$ if both b and ψ are zero, you don't have to worry about nasty infinities in the Bethe Free energy, and don't need to resort to annealing parameters. Another way to see this is that

$$b_{ij}(x_i, x_j) = \psi_{ij}(x_i, x_j) a_i(x_i) b_j(x_j) \quad (11)$$

where a_i and b_j are products of local evidences and messages coming into the nodes i and j , and that equation guarantees that if $\psi_{ij}(x_i, x_j) = 0$, then $b_{ij}(x_i, x_j) = 0$.

3. **How general is the pairwise Markov network?** In section 2 of (Yedidia *et al.*, 2000c), the undirected pairwise Markov network is introduced and said to be 'very general, as essentially any graphical model can be converted into this form'.

$$P(x_1, x_2, \dots, x_N) = \frac{1}{Z} \prod_{ij} \psi_{ij}(x_i, x_j) \prod_i \psi_i(x_i). \quad (12)$$

I am sceptical about this attitude. Is the pairwise Markov network really so general? In sparse graph codes, every parity node induces a potential relating at least three variables to each other – for example, if $z = x_1 + x_2$, then we have, as above,

$$\psi(x_1, x_2, z) = 1 + z + x_1 + x_2 \text{ mod } 2. \quad (13)$$

Typically we have parity check nodes with ten or so bits x participating in a single potential.

How would this conversion into pairwise Markov form work? I can see one way to do it: with a collection of 2^k additional binary variables, one for each setting of the k bits participating in the parity check. But that is horrible! (Is this the conversion that YTW had in mind?)

My feeling is that results that focus on the pairwise Markov network are too restricted. After all, the sum-product algorithm on a graph built from parity check potentials is very simple (Gallager, 1963; MacKay and Neal, 1996). I would find it helpful if the Bethe-energy enthusiasts could describe a version of the Bethe-free energy for a simple parity network like the one whose joint probability is defined by the potential (13). [No other potentials are needed for this example.]

JSY writes...

My answer to your question here takes several forms.

First, I should say that I agree with you that it is not always necessarily best to convert everything into a pairwise network. In (Yedidia *et al.*, 2000b), you'll see an example where we apply generalized belief propagation to a small parity check code. As we state in the draft, when we did this, we work directly in terms of the model written with higher-order interaction terms. In other words, we used tensorial ψ -interaction terms like $\psi_{ijklm}(x_i, x_j, x_k, x_l, x_m)$ which represented the 5 nodes x_i, x_j, x_k, x_l , and x_m being involved in a parity check. There is no restriction of the theory to pair-wise networks. In fact, I think that it is an advantage of generalized belief propagation that it applies so naturally to models with higher-order interactions.

But just to be clear: for the sum-product algorithm on any factor graph or Pearl’s algorithm on any directed graph there is a Bethe-like free energy that is being minimized. And an easy way to prove this is to convert these graphs to an equivalent pair-wise graph and use our claim 1.

For the specific example of the parity check that you give in equation (13), one can convert that into a pair-wise network as follows:

- (a) Introduce a new node w that can be in 8 different states corresponding to the possible states of x_1 , x_2 , and z . (We’ll call those 8 states 000, 001, 010, 011, 100, 101, 110, and 111).
- (b) Set up the ψ_{wz} , ψ_{wx_1} , and ψ_{wx_2} such that they equal 1 if the state of w matches the state of z , x_1 , or x_2 , respectively, and 0 if they don’t.
- (c) Set the evidence term $\psi_w = 1$ for the states 000, 011, 101, and 110, and 0 for the other states (to insure that the parity check is obeyed).

We describe this procedure in general for graphical model representations of parity check codes in section 2.2 of (Yedidia *et al.*, 2000a), and give the Bethe free energy for such models there.

4. **Undirectedness.** [This is not a fundamental question.] It bothers me that the Bethe/pairwise-Markov formulation uses an undirected graph, and consequently all the messages passed in the resulting algorithms are of one type only, whereas many expositions of the sum-product algorithm distinguish between two types of message – *probabilities* and *likelihoods* is one terminology for them. I’ve forgotten what I learned about the junction tree algorithm; maybe the junction tree algorithm has messages of one type only? So maybe the correspondence is especially direct between Bethe and JT?

JSY writes...

First of all, I realize you are not talking about directed graphs per se, but you will find in section 2.3 of (Yedidia *et al.*, 2000a) an explanation of how to convert a directed graph into an undirected pairwise graph.¹

What you are talking about, as I understand it, is that in most expositions of Gallager decoding, one has two kinds of messages: those from a check to a variable node about “what state you should be in”, and those from a variable node to a check about “what state I am in”. Both kinds of messages are binary if the variable nodes are binary. The messages in our papers, on the other hand, are always about “what state *you* should be in”, and if they are from a variable to a check, they range over the 2^k states of our check nodes.

These two formulations are mathematically equivalent. The conventional Gallager decoding amounts to defining messages which are

$$\mu_{\text{node}_i \rightarrow \text{check}_a}(x_i) = \alpha \psi_i(x_i) \prod_b m_{\text{check}_b \rightarrow \text{node}_i}(x_i) \quad (14)$$

These “ μ ” messages will just have as many states as the node i does (they will be binary).

For the sake of consistency with other kinds of models, we generally prefer to talk about messages

$$m_{\text{node}_i \rightarrow \text{check}_a}(x_a) = \alpha \sum_{x_i} \psi_{ia}(x_i, x_a) \mu_{\text{node}_i \rightarrow \text{check}_a}(x_i) \quad (15)$$

¹This material was also in the paper by Freeman and Weiss, to appear in IEEE Transactions on Information Theory, available at: <http://www.cs.berkeley.edu/~yweiss/maxprodieee2.pdf>

and our messages from nodes to checks range over 2^k states. But these methods are equivalent – it’s just a question of when you do the needed multiplication over ψ_{ia} and sum over x_i .

5. **The flipped-sign entropy terms.** Regular free energies and variational free energies contain *negative* entropy contributions (*e.g.*, $F = E - TS$), so that *large-entropy* probability distributions are favoured, all other things being equal. The Bethe free energy has *positive* entropies in it [the terms in $(q_i - 1)$]. Should I be worried about the possibility that these flipped-sign entropy terms might sometimes end up dominating things, giving silly answers in which the probability distribution has very *low* entropy?

JSY writes...

I do not think there is anything to worry about. The entropy terms with a flipped sign all involve single-node beliefs. If a single-node belief for a particular node has a concentrated probability distribution, the double-node beliefs involving that node must also have one by the marginalization constraints, and the entropy terms for those double-node beliefs will fight against the concentrated distribution because they still have the ordinary sign. It might look a little strange at first, but the Bethe entropy is actually the exact form on a tree, which might make you feel more comfortable about it.

ACKNOWLEDGEMENTS

David MacKay thanks Max Welling, Radford Neal, and Yee Whye Teh for helpful discussions.

References

- Gallager, R. G. (1963) *Low Density Parity Check Codes*. Number 21 in Research monograph series. Cambridge, Mass.: MIT Press.
- MacKay, D. J. C. (1995a) Free energy minimization algorithm for decoding and cryptanalysis. *Electronics Letters* **31** (6): 446–447.
- MacKay, D. J. C. (1995b) A free energy minimization framework for inference problems in modulo 2 arithmetic. In *Fast Software Encryption (Proceedings of 1994 K.U. Leuven Workshop on Cryptographic Algorithms)*, ed. by B. Preneel, number 1008 in Lecture Notes in Computer Science, pp. 179–195. Springer-Verlag.
- MacKay, D. J. C., and Neal, R. M. (1996) Near Shannon limit performance of low density parity check codes. *Electronics Letters* **32** (18): 1645–1646. Reprinted *Electronics Letters*, **33**(6):457–458, March 1997.
- Welling, M., and Teh, Y. W. (2001) Belief optimization for binary networks: A stable alternative to loopy belief propagation. Technical report, Gatsby Computational Neuroscience Unit. Check with Max for correct bibtex.
- Yedidia, J. S. (2000) An idiosyncratic journey beyond mean field theory. Technical report, Mitsubishi. MERL TR-2000-27.
- Yedidia, J. S., Freeman, W. T., and Weiss, Y. (2000a) Bethe free energy, kikuchi approximations and belief propagation algorithms. Technical report, Mitsubishi. MERL TR-2001-16.
- Yedidia, J. S., Freeman, W. T., and Weiss, Y. (2000b) Characterization of belief propagation and its generalizations. Technical report, Mitsubishi. MERL TR-2001-15.
- Yedidia, J. S., Freeman, W. T., and Weiss, Y. (2000c) Generalized belief propagation. Technical report, Mitsubishi. MERL TR-2000-26.
- Yuille, A. L., (2001) A double-loop algorithm to minimize the Bethe and Kikuchi free energies. Unpublished.