Belief Propagation and Statistical Physics

Payam Pakzad
EECS Department
University of California, Berkeley
payamp@eecs.berkeley.edu

Venkat Anantharam
EECS Department
University of California, Berkeley
ananth@eecs.berkeley.edu

Abstract — **It was shown recently in [1] that there is a close connection between the belief propagation algorithm and certain approximations to the variational free energy in statistical physics. Specifically, the fixed points of the belief propagation algorithm are shown to coincide with the stationary points of the Bethe's approximate free energy subject to consistency constraints. Bethe's approximation is known as a special case of a general class of approximations called Kikuchi free energy approximations. A general class of belief propagation algorithms was also introduced in [1], which attempts to find the stationary points of a general Kikuchi free energy functional.**

In this paper we first examine the physical significance of the 'free energy' functions, and review the general Kikuchi approximations using Möbius inversion formula. Next we define a general constrained minimization problem corresponding to the general Kikuchi approximation whose stationary points approximate marginals of a product function, and we specify a general class of local message passing algorithms along the edges of the Hasse diagram of the collection of Kikuchi regions, which attempt to solve that problem. We further give sufficient conditions under which a Kikuchi functional has a unique minimum, and/or closely approximates the exact free energy. These directly translate to conditions for the convergence and correctness of the belief propagation algorithm.

I. Free Energy and Kikuchi Approximation

Consider a system with N distinct and fixed sites. Attached to each site i is an elementary magnet which can take a spin value s_i from the set A of possible spins. Denote by $s = (s_1, \cdots, s_N)$ the configuration of the system. Then, under fundamental assumptions of thermal physics, it can be shown that if the system is in thermal equilibrium with a large *reservoir*, the probability that the system will be in a configuration s is given by the Boltzmann distribution:

$$P(s) = \frac{e^{-\varepsilon_s/\tau}}{Z} \tag{1}$$

where ε_s is the energy of the system at state configuration s, and τ is the temperature(see [2]). Z is called the *partition function* of the system and is defined by

$$Z = \sum_s e^{-\varepsilon_s/\tau} \tag{2}$$

The function

$$F = U - \tau S \tag{3}$$

is called the *Helmholtz (variational) free energy*, where $U = \sum_s P(s)\varepsilon_s$ is the average energy and $S = -\sum_s P(s)\ln(P(s))$ is the entropy of the system. We view F as a function of the distribution P.

It can be shown that the free energy F is minimized with the Boltzmann distribution, and at that point

$$F_0 := \min_{P(s)} F = -\tau \ln(Z) \tag{4}$$

Thermodynamical properties of the system can be derived if the free energy F_0 is known as a function of the temperature, e.g. $U = -\tau^2 \partial(F_0/\tau)/\partial\tau$ and $S = -\partial F_0/\partial\tau$. However, equation (4) does not prescribe a practical way to compute F_0 as it contains minimization over the exponentially large domain of distributions $P(s)$.

Suppose now that there is a collection R of the subsets of $\{1, \cdots, N\}$ such that the state energy function ε_s can be written as

$$\varepsilon_s = \sum_{r \in R} E_r(s_r) \tag{5}$$

for some set of energy functions E_r. Then the Boltzmann distribution is given by

$$P(s) = \frac{\prod_{r\in R} e^{-E_r(s_r)}}{Z} \tag{6}$$

and the average energy is

$$U = \sum_{r \in R} \sum_{s_r} P_r(s_r) E_r(s_r) \tag{7}$$

where $P_r(s_r)$ is the marginal of distribution $P(s)$. Note that the average energy is now only a function of the marginals $\{P_r(s_r)\}$.

Then from equation (4), we have

$$F_0 = \min_{\{P_r(s_r)\}} \left(U(\{P_r(s_r)\}) - \tau \tilde{S}(\{P_r(s_r)\}) \right) \tag{8}$$

where

$$\tilde{S}(\{P_r(s_r)\}) := \max_{P(s):\{P_r(s_r)\}} S(P(s))$$

is the maximum taken with respect to the joint distributions $P(s)$ that marginalize to a given collection $\{P_r(s_r)\}$. Note that not every collection $\{P_r(s_r)\}$ of probability functions can be marginals of a single distribution $P(s)$. In such cases we adopt the convention that $\tilde{S}(\{P_r(s_r)\}) = -\infty$.

Equation (8) is in the desirable form of replacing the minimization in (4) over complete distributions $P(s)$, by minimization over marginals $\{P_r(s_r)\}$. But we still need to estimate the entropy term $\tilde{S}(\{P_r(s_r)\})$. Following [3], we describe the estimation using the Möbius inversion formula.

Let $\hat{R}$ be the collection R of subsets of $C := \{1, \ldots, N\}$ together with the set C itself. Then $\hat{R}$ is a *poset* with the partial ordering of inclusion (see [4]). For each $r \in \hat{R}$ define the

regional entropy $S_r(P_r) := -\sum_{s_r} P_r(s_r)\ln(P_r(s_r))$, where, as before P_r's denote the marginals of a distribution $P(s)$. We wish to approximate the entropy $S = S_C$. Möbius dual functions $\bar{S}_r$'s are defined such that for each $t \in \hat{R}$,

$$S_t = \sum_{\substack{r\in\hat{R}\\ r\subseteq t}} \bar{S}_r \tag{9}$$

Then by the Möbius inversion formula

$$\bar{S}_r = \sum_{\substack{u\in\hat{R}\\ u\subseteq r}} S_u \mu(u,r) \tag{10}$$

Here the Möbius function $\mu(u,r)$ is defined for $u, r \in R, u \subseteq r$ by equations

$$\sum_{\substack{u\in R'\\ t\subseteq u\subseteq r}} \mu(u,r) = \delta_{tr} \tag{11}$$

Setting $r = C$ in (10) yields

$$S_C = S = -\sum_{r\in R} S_r \mu(r,C) + \bar{S}_C \tag{12}$$

If the term $\bar{S}_C$ can be ignored, we get the approximation

$$S \simeq -\sum_{r\in R} S_r \mu(r,C) \tag{13}$$

The Kikuchi's approximate free energy uses the above approximation of entropy in (3):

$$\begin{aligned} F_K(\{P_r(s_r)\}) &:= \big(U(\{P_r(s_r)\}) + \tau \sum_{r\in R} S_r(P_r(s_r))\mu(r,C)\big) \\ &= \sum_{r\in R}\sum_{s_r} (P_r(s_r)E_r(s_r) - \tau\mu(r,C)P_r(s_r)\ln(P_r(s_r))\big) \end{aligned} \tag{14}$$

This is identical to equation (35) in [1], where the 'overcounting factors' c_r in [1] are precisely the negatives of the corresponding Möbius factors $\mu(r,C)$, since from (11), $\mu(r,C) = -1$ for any maximal element r of R, and $\mu(u,C) = -1 - \sum_{r\supset u}\mu(r,C)$ for any $u \in R$. For the remainder of this paper we shall also use the shorthand $c_r = -\mu(r,C)$.

Free energy F_0 of equation (4) is estimated from the Kikuchi free energy (14) as

$$F_0 \simeq \min_{\{P_r(s_r)\}\in\mathcal{C}_R} F_K \tag{15}$$

where $\mathcal{C}_R$ is the constraint set of the pseudo-marginals that are locally consistent with respect to R:

$$\begin{aligned} &\mathcal{C}_R := \\ &\big\{\{P_r(s_r); r\in R\} \mid \forall t,u\in R,\, t\subset u,\, \sum_{s_{u\setminus t}} P_u(s_u) = P_t(s_t)\big\} \end{aligned} \tag{16}$$

Notice that here we are not only approximating the exact variational free energy F by its Kikuchi approximation F_K, but we are also replacing the minimization over the complete distributions $P(s)$ by minimization over consistent pseudo-marginals $\{P_r(s_r)\} \in \mathcal{C}_R$. The consistency constraint, however, is not enough in general to guarantee that a collection $\{P_r(s_r)\} \in \mathcal{C}_R$ are marginals of a single distribution function $P(s)$. In other words, the minimizing collection of pseudo-marginals $\{P_r(s_r)\}$ may not even belong to any real distribution.

II. A General Class of Constrained Minimization Problems

From the discussion in the previous section, the above method can be viewed as a way to estimate the marginals of a product function given by (6): given a set of local functions $\alpha_r(s_r) := e^{-E_r(s_r)}$, *if* the Kikuchi approximation F_K were exact, *and* the constrained pseudo-marginals corresponded to the marginals of a single distribution, then the minimizing P_r's of (15) would correspond *exactly* with the marginals of the Boltzmann distribution, $\prod_{r\in R}\alpha_r(s_r)/Z$; then, applying (15) with a good Kikuchi approximation F_K and set of constraints is expected to give a reasonable approximation of these marginals.

The collection R of regions effectively specifies both the Kikuchi approximation (14), and the constraint set $\mathcal{C}_R$. It is also evident that (15) as an approximation method can be applied for any given F_K and $\mathcal{C}_R$; better choices of R simply result in better approximations. Therefore we can define a general class of constrained minimization problems, which are specified by a poset R of regions, and local functions $\alpha_r(s_r)$ for each $r \in R$. (Note that although 'inclusion' is certainly the most natural partial ordering for R, the problem is well-defined for any arbitrary partial ordering. For conceptual comfort, however, we can assume that R is equipped with the partial ordering of inclusion for the remainder of this paper.)

A marginalization problem can be described by an objective 'marginalizable' function $\beta(s_1,\cdots,s_N)$ which is decomposable as a product of some functions $\alpha_r(s_r)$ for r's in a collection R of subsets of $C = \{1,\cdots,N\}$. We call the maximal elements of R the *basic clusters*; these are (typically) the largest non-decomposable factors of the objective function $\beta(s)$. For regions r, t, we include $r \cap t$ in R if we need to impose the constraint that the solutions P_r and P_t marginalize down to the same function $P_{r\cap t}$.

It is natural to represent poset R with its *Hasse diagram* G_R (see [4]). This is an undirected graph with an implied upward orientation, whose vertices are the elements of R and whose edges are the cover relations, and such that if $u \subset t$ then t is drawn "above" u. Therefore an edge exists between each $r \in R$ and its *immediate* parents, i.e. the minimal elements of R that properly contain r. As we shall see in the next section, there exist local message-passing algorithms along the edges of G_R, whose fixed points are solutions to the above constrained minimization problem.

It remains to specify which choices of R yield good approximations of the marginals. It certainly seems that minimization with more local consistency constraints on $\{P_r(s_r)\}$ should result in better approximations, since true marginals would satisfy all such constraints. Therefore one might conclude that including more subregions in R should improve the approximation (at the expense of increasing the complexity of G_R and its corresponding algorithm).

It is natural to require that all the regions in R which contain a given index $i \in \{1,\cdots,N\}$ be connected in the Hasse diagram G_R (Condition (**A**1)). This will ensure that the beliefs $P_r(s_r)$ at all the regions r which contain index i will be consistent at the level of variable s_i. Based on this, we can devise condition (**A**n), which requires that for each $m \le n$, all the regions containing a given m-tuple of indices $\{i_1,\cdots,i_m\}$ be connected.

Inspired by [5], one might insist that acceptable approximations of the entropy term (13) are those in which each variable

s_i appears the same number of times on the two sides of the equality sign, i.e.

$$\sum_{r:i\in r} c_r = 1 \qquad \text{for each } i = 1, \cdots, N \qquad (\mathbf{B}1)$$

We can extend this condition also, as follows:

$$\sum_{r:s\subseteq r} c_r = 1 \qquad \text{for each } s \subset \{1, \cdots, N\},\ |s| \le n \qquad (\mathbf{B}n)$$

These conditions are expected to give progressively better approximate solutions. It is noteworthy that the original Kikuchi collection of regions as defined in [3] and [1] was required to be closed under intersection. It can be shown that any collection of regions R which is closed under intersection satisfies $(\mathbf{A}n)$ and $(\mathbf{B}n)$ for all n.

The special case when the Hasse diagram G_R has depth 2, i.e. there are no distinct $r, s, t \in R$ such that $r \subset s \subset t$, is called the Bethe case. In this case G_R can be thought of as a hypergraph in which the basic clusters are the vertices and other regions are the hyperedges. If we insist that the basic clusters be pairs $\{i, j\}$ of indices for $i, j \in \{1, \cdots, N\}$ as assumed in [1], we will in fact have a poset R of depth 2 which is closed under intersection. But as discussed here, the restriction on the size of clusters is in fact unnecessary, as we allow for R not to be closed under intersection.

On the other hand, [5] considers only the case when the aforementioned 'hypergraph' view of G_R is a graph (i.e. the minimal elements of R are covered by at most two basic clusters, so the hyperedges are in fact edges). The 'junction graph' condition given in [5] is simply the intersection of conditions $(\mathbf{A}1)$ and $(\mathbf{B}1)$ above. (It can be shown that the 'junction graph' condition does *not* imply $(\mathbf{A}2)$ or $(\mathbf{B}2)$.)

It can be verified that neither of definitions above from [1] and [5] includes the other and, as shown here, our definition includes both of them as special cases, and in that sense our formulation is more general than both [1] and [5].

III. Lagrange Multipliers and Iterative Solutions

Lagrange's method can be used to solve the constrained minimization problem (15). Taking $\tau = 1$, and using $c_r = -\mu(r, C)$ and $\alpha_r(s_r) = e^{-E_r(s_r)}$ the Lagrangian becomes:

$$\begin{aligned}\mathcal{L} \;:=\; & \sum_{r\in R}\sum_{s_r}\big(-P_r(s_r)\ln(\alpha_r(s_r)) + c_r P_r(s_r)\ln(P_r(s_r))\big) \\ & + \sum_{r\in R}\sum_{t\prec r}\sum_{s_t}\lambda_{rt}(s_t)\big(P_t(s_t) - \sum_{s_{r\setminus t}} P_r(s_r)\big) \\ & + \sum_{r\in R}\kappa_r\big(\sum_{s_r} P_r(s_r) - 1\big) \qquad (17)\end{aligned}$$

where coefficients $\lambda_{rt}(s_t)$ enforce consistency constraints, and coefficients κ_r enforce normalization constraints, and by $t \prec r$ in the second line we mean that r covers t, i.e. t is maximal in the subset of elements of R that are properly contained in r. Note that we need only define λ_{rt} for pairs $r, t \in R$ with $t \prec r$, i.e. along the edges of G_R.

Setting partial derivative $\partial\mathcal{L}/\partial P_r(s_r) = 0$ for each $r \in R$ gives an equation for $P_r(s_r)$ in terms of λ_{ur}'s and λ_{rt}'s. The consistency constraints give update rules for each λ_{rt} in terms of other λ factors. Once a set of messages m_{rt} (from r to t, with $t \prec r$) has been defined in terms of the Lagrange factors λ_{rt}'s, these update rules define an iterative algorithm whose fixed points are the stationary points of the given constrained minimization problem.

One particularly nice such algorithm is the 'generalized belief propagation' discussed in [1] which defines the messages so that belief $P_r(s_r)$ depends only on the outside messages to a subregion of r. Belief propagation algorithm can also be seen as one such iterative algorithm in the Bethe case (see [1] and [5]).

The Kikuchi free energy (14) is bounded below and hence the constrained minimization problem (15) always has a global minimum. Therefore the above message passing algorithms always possess at least one fixed point (see [6] for an algorithm that is guaranteed to find a minimum of F_K).

The following result gives sufficient conditions on R for the problem (15) to have precisely one minimum:

Theorem 1. *The Kikuchi free energy functional* (14) *is convex over the set of consistency constraints imposed by a collection of regions R (and hence the constrained minimization problem has a unique solution) if the overcounting factors $c_r, r \in R$ satisfy:*

$$\forall S \subset R,\ \sum_{t\in S} c_t + \sum_{\substack{r\in R\setminus S:\\ \exists t\in S, t\subset r}} c_r \ge 0 \qquad (18)$$

In words, for any subset S of R, the sum of overcounting factors of elements of S and all their ancestors in R must be nonnegative.

Remember that in the Bethe case, for basic clusters $r \in R$, $c_r = 1$, and for the other regions $t \in R$, $c_t = 1 -$ (no. of covers of t). Thus we have

Corollary 2. *In the Bethe case, the constrained minimization problem* (15) *has a unique solution if the graphical representation G_R of R has at most one loop.*

Acknowledgments

This work was supported by grants from (ONR/MURI) N00014-1-0637, (NSF) SBR-9873086, (DARPA) F30602-00-2-0538, California Micro Program, Texas Instruments, Marvel Inc. and ST MicroElectronics.

References

[1] J.S. Yedidia, W.T. Freeman and Y. Weiss, *"Bethe Free Energy, Kikuchi Approximations, and Belief Propagation Algorithms,"*, MERL T.R., 2000.

[2] C. Kittel and H. Kroemer, *Thermal Physics*, New York: W.H. Freeman & Co., 1980.

[3] T. Morita, *"Formal Structure of the Cluster Variation Method,"* Prog. Theor. Phys. Supp. No. 115, 1994, pp.27–39.

[4] R.P. Stanley, *Enumerative Combinatorics, volume I*, Monterey, CA: Wadsworth & Brooks/Cole, 1986.

[5] S.M. Aji and R.J. McEliece, *"The Generalized Distributive Law and Free Energy Minimization,"* Allerton Conference Presentation, Oct. 2001.

[6] A.L. Yuille, *"A Double-Loop Algorithm to Minimize the Bethe and Kikuchi Free Energies,"* to appear in Neural Computation.

Estimators based on belief propagation have received a lot of attention due to their low complexity and good performance [7]â€“[9]. Citing wisdom in statistical physics and neural networks, several works suggest that the large-system performance of BP is described by the meta-stable solutions of the xed-point equations of [5], [6].Â The same formula (11) is claimed in [6] for general (dense). linear systems without rigorous justication. B. Optimality of Belief Propagation. Consider the bipartite graph (Fig. 1) which describes the linear system (1). An iterative BP estimator can be devised based on the graph, which essentially updates the posterior distribution (or â€œbeliefâ€) for each Xk conditioned on the observations within the reach of the local subtree Y (t) with Xk as the root [19], [20]. Belief propagation is a widely used message passing method for the solution of probabilistic models on networks such as epidemic models, spin models, and Bayesian graphical models, but it suffers from the serious shortcoming that it works poorly in the common case of networks that contain short loops.Â Perhaps the best known example of generalized belief propagation, at least within the statistical physics community, is the cluster variational method, in which the regions are defined so as to be closed under the intersection operation (24), and the resulting free energy is called the Kikuchi free energy (50). apport de recherche. Belief Propagation and Bethe approximation for Trac Prediction. Cyril Furtlehnerâˆ—, Jean-Marc Lasgouttesâ€ , Arnaud de La Fortelleâ€¡â€ . Th`emes NUM et COG â€” Syst`emes numÂ´eriques et Syst`emes cognitifs Projets Imara et Tao. Rapport de recherche n ???? â€” Mars 2007 â€” 29 pages. Abstract: We dene and study an inference algorithm based on â€œbelief propagationâ€ (BP) and the Bethe approximation.Â 2 Trac description and statistical physics. The graph onto which we apply the belief propagation procedure is made of space-. time vertices that encode both a location (road link) and a time (discretized on a few minutes scale).