Handout Probabilistic Graphical Models: Bethe Free Energy and Loopy Belief Propagation

Volkan Cevher, Matthias Seeger

Abstract

Here is the proof of [4] in the notation of our lecture. For full generality, have a look at their paper and at [3].

A factor graph induces the distribution

$$P(x_1, \ldots, x_n) = Z^{-1} \prod_{j=1}^{J} \Phi_j(\boldsymbol{x}_{C_j}),$$

where $C_j \subset \{1, \ldots, n\}$. For simplicity, all potentials are positive functions here. For the general case, see [4]. Loopy belief propagation (LBP) defines messages (positive functions over variables), which are initially all uniform, and iterates the following equations in some ordering until convergence.

$$M_{i\to j}(x_i) \propto \prod_{j' \in \mathcal{N}_i \setminus j} M_{j'\to i}(x_i), \quad M_{j\to i}(x_i) \propto \sum_{\boldsymbol{x}_{C_j \setminus i}} \Phi_j(\boldsymbol{x}_{C_j}) \prod_{i' \in C_j \setminus i} M_{i'\to j}(x_{i'}). \tag{1}$$

Recall that $\mathcal{N}_i = \{j \,|\, i \in C_j\}$: the neighbouring factor nodes of variable node i. In case this convergences, marginals are approximated by

$$\mu_i(x_i) \propto \prod_{j \in \mathcal{N}_i} M_{j\to i}(x_i), \quad \mu_j(\boldsymbol{x}_{C_j}) \propto \Phi_j(\boldsymbol{x}_{C_j}) \prod_{i \in C_j} M_{i\to j}(x_i). \tag{2}$$

I will use i as general index over variables, j over factors. If ranges of sums or products are not given, they run over all permitted values.

In general, LBP may not converge. If it does, $\mu_i(x_i)$ may not be close to the true marginal $P(x_i)$. In fact, the set of pseudomarginals $\boldsymbol{\mu}$ may not even be realizable, in the sense that they are not legal marginals of any joint distribution. For a general graph, even this test is hard.

The hope is, of course, that convergence occurs to good marginal approximations. Because this algorithm tends to perform amazingly well on many problems, there is an ever growing literature on convergence analyses, computable error bounds, modifications such as convexification, different convergent algorithms, and so on. The online article [2] gives a wide current overview.

The Bethe free energy is

$$\mathcal{F}_{\text{Bethe}} = \sum_j \mathrm{E}_{\mu_j}[-\log \Phi_j(\boldsymbol{x}_{C_j})] - \sum_j \mathrm{H}[\mu_j(\boldsymbol{x}_{C_j})] + \sum_i (n_i - 1)\mathrm{H}[\mu_i(x_i)], \quad n_i = |\mathcal{N}_i|. \tag{3}$$

The constraint set for $\boldsymbol{\mu}$ is the local marginalization polytope $\mathcal{M}_{\text{local}}$, defined by

$$\mu_j(\boldsymbol{x}_{C_j}) \geq 0, \quad \mu_i(x_i) \geq 0, \quad \sum_{x_i} \mu_i(x_i) = 1, \quad \sum_{\boldsymbol{x}_{C_j \setminus i}} \mu_j(\boldsymbol{x}_{C_j}) = \mu_i(x_i), \tag{4}$$

with all-quantification over all variables not summed over. The claim is that a fixed point of LBP, where no message update (1) leads to changes anymore, is a stationary (saddle) point of $\mathcal{F}_{\text{Bethe}}$, subject to $\boldsymbol{\mu} \in \mathcal{M}_{\text{local}}$.

To show that, I use Lagrange duality theory. If you don't know what that is, you better read about it, say in [1]. It is a cornerstone of convex programming, and of relaxations to non-convex problems, and can be used to characterize saddle points (as I do here). I'll be lazy and dualize equality constraints only, leaving the nonnegativity constraints "un-dualized". Under the assumption of positive potentials, this goes through (see [2, Remark 4.1]). I need one multiplier λ_i for each $\sum_{x_i} \mu_i(x_i) = 1$, and multipliers $\lambda_{i \to j}(x_i)$ for each j, $i \in C_j$, and x_i (for $\sum_{\boldsymbol{x}_{C_j \setminus i}} \mu_j(\boldsymbol{x}_{C_j}) = \mu_i(x_i)$). Recall from the lecture that "(x_i)" is nothing more than an index into a vector. I could also write λ_{i,j,x_i} or $\lambda_{i,j,k}$, but the functional notation is simpler. To make things simpler, I dualize the implied constraints $\sum_{\boldsymbol{x}_{C_j}} \mu_j(\boldsymbol{x}_{C_j}) = 1$ as well, using multipliers λ_j. The Lagrangian is

$$\begin{aligned}
\mathcal{L} = & \sum_j \mathrm{E}_{\mu_j}[-\log \Phi_j(\boldsymbol{x}_{C_j})] - \sum_j \mathrm{H}[\mu_j(\boldsymbol{x}_{C_j})] + \sum_i (n_i - 1)\mathrm{H}[\mu_i(x_i)] \\
& - \sum_i \sum_{j \in \mathcal{N}_i} \sum_{x_i} \lambda_{i \to j}(x_i) \left(\sum\nolimits_{\boldsymbol{x}_{C_j \setminus i}} \mu_j(\boldsymbol{x}_{C_j}) - \mu_i(x_i) \right) - \sum_i \lambda_i \left(1 - \sum\nolimits_{x_i} \mu_i(x_i) \right) \\
& - \sum_j \lambda_j \left(1 - \sum\nolimits_{\boldsymbol{x}_{C_j}} \mu_j(\boldsymbol{x}_{C_j}) \right).
\end{aligned}$$

It is a function of $\boldsymbol{\mu}$ (which in the Lagrange dual problem does not have to follow the equality constraints) and $\boldsymbol{\lambda}$, and the dual problem is $\min_{\boldsymbol{\mu}} \max_{\boldsymbol{\lambda}} \mathcal{L}$, subject to $\boldsymbol{\mu} \succeq \mathbf{0}$. A saddle point is a pair $\boldsymbol{\mu}_*$, $\boldsymbol{\lambda}_*$ such that $\nabla_{\boldsymbol{\mu}_*} \mathcal{L} = \mathbf{0}$ and $\nabla_{\boldsymbol{\lambda}_*} \mathcal{L} = \mathbf{0}$.

I have to show that at a fixed point of LBP, these stationary equations are met. First, as always with Langrange duals, $\nabla_{\boldsymbol{\lambda}} \mathcal{L} = \mathbf{0}$ is equivalent to $\boldsymbol{\mu} \in \mathcal{M}_{\text{local}}$ (recall that $\boldsymbol{\mu} \succeq \mathbf{0}$ is always true). Also, since the Lagrangian is linear in $\boldsymbol{\lambda}$, I can reparameterize these multiplier variables linearly, without modifying their key property (namely, that $\nabla_{\boldsymbol{\lambda}} \mathcal{L} = \mathbf{0}$ implies $\boldsymbol{\mu} \in \mathcal{M}_{\text{local}}$). I'll now play around with $\nabla_{\boldsymbol{\mu}} \mathcal{L} = \mathbf{0}$ and see where I get at, always ensuring $\boldsymbol{\mu} \in \mathcal{M}_{\text{local}}$. Of course, if everything works out, I will arrive at the message passing and pseudomarginal equations, but in order to not get confused, forget about them at the moment.

Since $d\mathrm{H}[P(x)] = -d \sum_x P(x) \log P(x) = -(\log P(x) + 1)(dP(x))$, we have

$$\frac{\partial \mathcal{L}}{\partial \mu_j(\boldsymbol{x}_{C_j})} = -\log \Phi_j(\boldsymbol{x}_{C_j}) + \mu_j(\boldsymbol{x}_{C_j}) + 1 - \sum_{i \in C_j} \lambda_{i \to j}(x_i) + \lambda_j = 0.$$

Here, I used that $\sum_i \sum_{j \in \mathcal{N}_i} = \sum_{i,j} \mathrm{I}_{\{j \in \mathcal{N}_i\}} = \sum_{j,i} \mathrm{I}_{\{i \in C_j\}}$, in general a good way to deal with such sums. Lumping constants into λ_j, then *defining* $M_{i \to j}(x_i) := e^{\lambda_{i \to j}(x_i)}$, this becomes

$$\mu_j(\boldsymbol{x}_{C_j}) = e^{-\lambda_j} \Phi_j(\boldsymbol{x}_{C_j}) \prod_{i \in C_j} M_{i \to j}(x_i). \tag{5}$$

Next, for i with $n_i > 1$,

$$\frac{\partial \mathcal{L}}{\partial \mu_i(x_i)} = -(n_i - 1)(\mu_i(x_i) + 1) + \sum_{j \in \mathcal{N}_i} \lambda_{i \to j}(x_i) + \lambda_i = 0. \tag{6}$$

I define some auxiliary variables $\tilde{\lambda}_{j \to i}(x_i)$ by $\lambda_{i \to j}(x_i) = \sum_{j' \in \mathcal{N}_i \setminus j} \tilde{\lambda}_{j' \to i}(x_i)$, moreover $M_{j \to i}(x_i) := e^{\tilde{\lambda}_{j \to i}(x_i)}$. This means that

$$M_{i \to j}(x_i) = \prod_{j' \in \mathcal{N}_i \setminus j} M_{j' \to i}(x_i) \tag{7}$$

Moreover,

$$\sum_{j \in \mathcal{N}_i} \lambda_{i \to j}(x_i) = \sum_{j, j' \in \mathcal{N}_i} \mathrm{I}_{\{j \neq j'\}} \tilde{\lambda}_{j' \to i}(x_i) = (n_i - 1) \sum_{j \in \mathcal{N}_i} \tilde{\lambda}_{j \to i}(x_i),$$

and after suitable linear transformation of λ_i, (6) becomes

$$\mu_i(x_i) = e^{-\lambda_i} \prod_{j \in \mathcal{N}_i} M_{j \to i}(x_i), \quad n_i > 1. \tag{8}$$

Now, we need to enforce $\boldsymbol{\mu} \in \mathcal{M}_{\text{local}}$. The normalization constraints are dealt with by setting λ_i and λ_j accordingly. Let us plug (5) and (8) into the consistency constraints (for i with $n_i > 1$):

$$e^{-\lambda_j} \sum_{\boldsymbol{x}_{C_j \setminus i}} \Phi_j(\boldsymbol{x}_{C_j}) \prod_{i' \in C_j} M_{i' \to j}(x_{i'}) = e^{-\lambda_i} \prod_{j' \in \mathcal{N}_i} M_{j' \to i}(x_i) = e^{-\lambda_i} M_{j \to i}(x_i) M_{i \to j}(x_i),$$

where I used (7). Dividing both sides by $M_{i \to j}(x_i)$, we have

$$e^{\lambda_i - \lambda_j} \sum_{\boldsymbol{x}_{C_j \setminus i}} \Phi_j(\boldsymbol{x}_{C_j}) \prod_{i' \in C_j \setminus i} M_{i' \to j}(x_{i'}) = M_{j \to i}(x_i). \tag{9}$$

Summing up, I have shown that if (5), (7), (8), and (9) hold, then $\boldsymbol{\mu} \in \mathcal{M}_{\text{local}}$ and $\nabla_{\boldsymbol{\mu}} \mathcal{L} = \mathbf{0}$. Together, they form a sufficient condition for a stationary point of the Lagrangian, therefore for a saddle point of the constrained Bethe problem. Obviously, at a fixed point of LBP, these conditions all hold (because message passing does not lead to any changes). You might object, because messages are defined only up to normalization, while (9) seems to demand a particular normalization. But at a LBP fixed point, you can just renormalize all messages at will, and you will stay there. At this point, the constraint $\boldsymbol{\mu} \in \mathcal{M}_{\text{local}}$ prescribes the normalization of (9).

References

[1] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[2] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008.

[3] J. Yedidia, W. Freeman, and Y. Weiss. Understanding belief propagation and its generalizations. Technical Report TR-2001-22, MERL, Cambridge, Massachusetts, January 2002.

[4] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Generalized belief propagation. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 689–695. MIT Press, 2001.