Compact Belief Propagation for Protein Folding

Jian Peng, Tamir Hazan, David McAllester, Raquel Urtasun
TTI Chicago

Many applications such as protein folding and stereo vision can be described using graphical models, where interactions between variables correspond to hyperedges in the graph. In these applications one is usually interested in computing the best configuration of the variables, usually referred to as the maximum a-posteriori (MAP) assignment. When dealing with discrete variables graphical models encode the complexity of NP-hard problems. When the variables are continuous only local minimum can be recovered in general.

Existing approaches for graphical models over continuous variables are Gaussian belief propagation for quadratic programs [1] as well as non-parametric belief propagation [6] and particle belief propagation [3] for general programs. The latter two frameworks lack optimality guarantees, i.e. even if these algorithms converge there is no guarantee that their solution recover the MAP assignment. Recently, optimality guarantees were given for a family of convex belief propagation algorithms for discrete sets [5, 2]. However, some of the optimality guarantees do not hold for general programs.

In this paper we propose to use duality to extend convex max-product to deal with compact spaces, i.e., mixtures of discrete and continuous bounded variables. Our algorithm is guaranteed to converge, but more importantly we derive the theoretical conditions under which it is guaranteed to obtain the optimal MAP assignment. The resulting messages are continuous functions. We derive different message representations and demonstrate the effectiveness of our approach in the task of protein folding and show that our approach significantly outperforms particle max-product and performs comparable to the state-of-the-art.

We are interested in graphical models whose variables are either discrete and bounded continuous. These graphical models typically consider functions defined on a single variable which correspond to the vertices in the graph, i.e., $\theta_i(x_i)$, and functions of the form $\theta_\alpha(x_\alpha)$ which are defined over subsets of variables $\alpha \subset \{1, .., n\}$ and correspond to the graph hyperedges. A general graphical program has the form:

$$\operatorname*{argmax}_{x_1,\ldots,x_n} \sum_{i \in V} \theta_i(x_i) + \sum_{\alpha \in E} \theta_\alpha(x_\alpha). \tag{1}$$

The main problem when using the convex max-product for general programs is recovering the MAP assignment from the algorithm's output. Here, we use the duality between continuous functions and regular Borel measures over compact spaces to show that convex max-product can recover the optimal MAP assignment when dealing with compact spaces. This is not very restricted since a wide range of applications can be solved in such spaces. In order to describe the duality framework for optimizing (1) we first transform this program into a continuous linear program with non-convex constraints, and then derive its dual. We show that the dual optimal solution can be obtained using convex max-product and derive the sufficient conditions for optimality, as well as for recovering the MAP assignment. Let K_i be the compact set of x_i and let K_α be the cartesian product of the compact sets K_i over $i \in N(\alpha)$. The objective in (1) can be described by a linear function of the form $\sum_\alpha \langle \theta_\alpha, \cdot \rangle + \sum_i \langle \theta_i, \cdot \rangle$. The continuous linear functions over θ_α, θ_i are identified with the set of regular Borel measures over the compact sets K_α and K_i. We use point mass measures, i.e., probability measures that concentrate all their weight on a single point, to formulate the program in (1) as

$$\max_{\delta_i, \delta_\alpha} \quad \sum_\alpha \langle \theta_\alpha, \delta_\alpha \rangle + \sum_i \langle \theta_i, \delta_i \rangle \tag{2}$$

$$\text{subject to:} \qquad \forall i, x_i, \alpha \in N(i), \int_{x_\alpha \setminus x_i} \delta_\alpha(x_\alpha) = \delta_i(x_i) \qquad \delta_i, \delta_\alpha \text{ are point mass probability measures.}$$

We have traded the computational complexity of the objective in (1) with the one of the feasible set in (2), thus the computational complexity remains high. As the set of all point mass measures is not convex, this linear program cannot be efficiently solved in general. We convexify the program by constructing its dual: $q(\lambda_{i\to\alpha}) = \sum_\alpha \max_{x_\alpha \in K_\alpha} \left\{ \theta_\alpha(x_\alpha) + \sum_{i \in N(\alpha)} \lambda_{i\to\alpha}(x_i) \right\} + \sum_i \max_{x_i \in K_i} \left\{ \theta_i(x_i) - \sum_{\alpha \in N(i)} \lambda_{i\to\alpha}(x_i) \right\}$ and derive optimality conditions for the recovery of the MAP assignment in (1).

Protein	length	#tem	MODELLER	PBP	Ours
T0437	99	1	**61.6**	32.7	60.9
T0451	133	2	64.3	27.8	**67.1**
T0464	89	1	42.3	28.7	**44.1**
T0471	133	2	57.1	31.2	**57.9**
T0473	68	1	**90.4**	42.1	88.9
T0522	134	2	**94.4**	32.4	93.2
T0562	123	1	33.5	25.5	**35.7**
T0574	126	2	55.1	31.2	**57.8**
T0579	124	1	42.9	26.8	**44.1**
T0592	144	2	**73.5**	25.8	72.3
T0606	123	1	**69.5**	32.6	69.1
T0610	186	1	**69.8**	28.7	66.2
T0622	138	2	60.5	30.5	**62.5**
T0630	132	1	54.4	22.8	**56.2**
1ctf	68	1	73.1	29.3	**74.7**
4icb	76	2	49.2	24.6	**54.4**
2cro	65	1	**84.2**	36.2	83.6
1fc2	43	1	64.2	27.9	**67.8**
2gb1	56	1	86.7	40.1	**87.0**
1enh	54	1	**88.2**	43.0	87.9

Table 1: **Comparison to the baselines** on 20 proteins ranging from 43 to 186 nodes and 788 to 16833 cliques.

Claim 1 *Let $X_i^* = argmax_{x_i}\{\theta_i(x_i) - \sum_{\alpha \in N(i)} \lambda_{i\to\alpha}(x_i)\}$ and $X_\alpha^* = argmax_{x_\alpha}\{\theta_\alpha(x_\alpha) + \sum_{i \in N(\alpha)} \lambda_{i\to\alpha}(x_i)\}$. If there exist probability measures b_α, b_i whose supports are contained in X_α^*, X_i^* and they agree on their marginals, then the continuous functions $\lambda_{i\to\alpha}(x_i)$ are dual optimal. If b_α, b_i are point mass distribution then they point towards an optimal MAP assignment. In particular, if the functions $\theta_i(x_i) - \sum_{\alpha \in N(i)} \lambda_{i\to\alpha}(x_i)$ have no ties then $x_1^*, ..., x_n^*$ is the MAP assignment.*

The main computational difficulty in applying message-passing to compact Euclidean sets is that the messages are continuous functions over K_i. We consider two types of message representations: The first one is inspired by particle approaches, where the messages are represented by their value on a dynamic set of points. The second representation is a multi-grid method where the space is partitioned into disjoint sets, on which the messages have a piecewise constant representation. While the former is very efficient only the latter conserves the theoretical guarantees.

We demonstrate the effectiveness of our approach in the task of template-based protein folding which can be formulated as a hybrid problem where continuous variables represent 3D locations of the C_α atoms in the backbone, while discrete variables represent the choice of template. As shown in Table 1, our approach significantly outperforms particle max-product. We also show that while using a very simple energy function based on distance constraints between pairs of C_α atoms, we perform comparably to the state-of-the-art, i.e., MODELLER [4], which takes into account a larger set of physical constrains derived from prior knowledge.

References

[1] D. Bickson. Gaussian Belief Propagation: Theory and Aplication. *Arxiv preprint arXiv:0811.2518*, 2008.

[2] T. Hazan and A. Shashua. Norm-Product Belief Propagation: Primal-Dual Message-Passing for Approximate Inference. *Arxiv preprint arXiv:0903.3127*, 2009.

[3] A. Ihler and D. McAllester. Particle belief propagation. In *AISTATS*, 2009.

[4] A. Sali and T. L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *Journal of molecular biology*, 234(3), December 1993.

[5] D. Sontag and T. Jaakkola. Tree block coordinate descent for MAP in graphical models. In *AISTATS*, 2009.

[6] E.B. Sudderth, A.T. Ihler, M. Isard, W.T. Freeman, and A.S. Willsky. Nonparametric belief propagation. *CACM*, 53(10), 2010.

Topic: Graphical models, Computational biology
Preference: Poster