Message Passing for Collective Graphical Models

Tao Sun and **Daniel Sheldon**
School of Computer Science
University of Massachusetts Amherst
Amherst, MA 01003
{taosun,sheldon}@cs.umass.edu

Akshat Kumar
IBM Research
New Delhi 110070
India
akshat.kumar@gmail.com

Abstract

Collective graphical models (CGMs) are a formalism for inference and learning with aggregate data that are motivated by a model for bird migration. We highlight a close connection between approximate MAP inference in CGMs and *marginal inference* in standard graphical models. The connection leads us to derive a novel Belief Propagation (BP)-style algorithm for collective graphical models. The algorithm is a strict generalization of BP, and is much more efficient than previous approaches to inference in CGMs. We demonstrate its performance on both synthetic and real datasets concerning the bird migration problem.

1 Introduction

In an influential paper, Yedidia, Freeman, and Weiss (2000) showed that the loopy Belief Propagation (BP) algorithm for marginal inference in graphical models can be understood as a fixed-point iteration that attempts to satisfy the first-order optimality conditions of the Bethe free energy, which approximates the true variational formulation of marginal inference. The result shed considerable light on the nature of BP and led to many new ideas for approximate variational inference. An interesting aspect of this result is that it began with a simple and well-known algorithm (loopy BP) and developed the theory to retrofit an explanation in terms of the Bethe free energy.

In this paper, we note a striking similarity between the Bethe free energy and the objective function for approximate MAP inference in CGMs (Sheldon et al., 2013), and then follow reasoning similar to that of Yedidia et al. but in the reverse direction to guide us to a novel message-passing algorithm for CGMs. The resulting algorithm has the interesting property that message updates are identical to BP, *with the exception that edge potentials change in each step* based on the gradient of the "evidence terms" that are present in the CGM objective but not in the Bethe free energy. The algorithm can be seen as a strict generalization of BP to deal with the presence of these additional non-linear terms.

The new algorithm has great practical benefits. We show experimentally that, by exploiting the graph structure, message passing solves the approximate MAP optimization problem much faster than generic solvers, and scales significantly better than any previous approach for inference in CGMs. For the problem of modeling bird migration, the new algorithm allows us to move from toy-sized problems to realistic models of migration in the eastern US. We present preliminary results that use this algorithm to infer the migration routes of birds from data collected by volunteer birdwatchers through the eBird citizen science project (Sullivan et al., 2009).

2 Collective Graphical Models

Sheldon and Dietterich (2011) introduced *collective graphical models* (CGMs) to model problems of learning and inference with noisy aggregate data. The motivating application is the problem of modeling bird migration from eBird data (Sheldon et al., 2008, 2013; Sheldon, 2009). CGMs may also be applied in other areas such as social science where individual data is difficult to collect but aggregate data is readily available.

The CGM generative model. CGMs compactly describe the distribution of the aggregate statistics of a population sampled independently from a discrete graphical model. Let $G = (V, E)$ be an undirected graph, and let $p(\mathbf{x}) = \Pr(X = \mathbf{x}) = \prod_{(i,j)\in E} \phi_{ij}(x_i, x_j)$ be a graphical model over G, i.e., a distribution over the discrete random vector $X = (X_1, \ldots, X_{|V|})$. Now, consider a population $X^{(1)}, \ldots, X^{(M)}$ of random vectors sampled independently from the graphical model. Define the contingency tables $\mathbf{n}_i = \{n_i(x_i)\}$ over nodes of the model and $\mathbf{n}_{ij} = \{n_{ij}(x_i, x_j)\}$ over edges of the model, whose entries count the number of times particular variable settings occur in the population:

$$n_i(x_i) = \sum_{m=1}^{M} \mathbb{1}[X_i^{(m)} = x_i], \quad n_{ij}(x_i, x_j) = \sum_{m=1}^{M} \mathbb{1}[X_i^{(m)} = x_i, X_j^{(m)} = x_j].$$

Note that these tables are random variables that depend on the entire population, and, for tree-structured models, which are the focus of this paper, the edge tables are the sufficient statistics of the model. For general models, all of the results in this section can be generalized to junction trees, with the usual blowup in space and running-time depending on the clique-width of the junction tree. In CGMs, one makes noisy observations of the sufficient statistics or their subtables in the form of a vector $\mathbf{y}$, and the goal is to compute the posterior distribution $p(\mathbf{n} \mid \mathbf{y}) \propto p(\mathbf{n})p(\mathbf{y} \mid \mathbf{n})$, where $\mathbf{n} = \{\mathbf{n}_i, \mathbf{n}_{ij}\}$. For efficient inference, we require that $p(\mathbf{y} \mid \mathbf{n})$ is log-concave in $\mathbf{n}$. In this work, we further assume that only node tables are observed and the entries $y_i(x_i)$ of $\mathbf{y}$ are generated independently from the corresponding node table entries $n_i(x_i)$ according to a univariate log-concave noise process $p(y \mid n)$, though these assumptions can be relaxed. We refer to the log-likelihood function $\ell_{i,x_i}(n_i(x_i)) = \log p(y_i(x_i) \mid n_i(x_i))$, a univariate concave function of $n_i(x_i)$, as the CGM *evidence* function.

Example. For modeling bird migration, assume that $X = (X_1, \ldots, X_T)$ is the sequence of discrete locations (e.g. map grid cells) visited by an individual bird, and that the graphical model $p(\mathbf{x}) = \prod_{t=1}^{T-1} \phi(x_t, x_{t+1})$ is a Markov chain governing the migration of the individual. M birds of species S independently migrate from location to location according to the Markov chain. The node-table entries $n_t(x_t)$ indicate how many birds are in location x_t at time t. The edge-table entries $n_{t,t+1}(x_t, x_{t+1})$ indicate how many birds move from location x_t to location x_{t+1} at time t. A reasonable model for eBird data is that the number of birds of species S counted by a birdwatcher is a Poisson random variable with mean proportional to the true number of birds plus some background rate, or: $y_i(x_i)|n_i(x_i) \sim \text{Pois}(\alpha n_i(x_i) + \alpha_o)$. Given only the noisy eBird counts and the prior specification of the Markov chain, the goal is to answer queries about the distribution $p(\mathbf{n} \mid \mathbf{y})$ to inform us about migratory transitions made by the population. Because the vector $\mathbf{n}$ includes the sufficient statistics, these queries also provide all the relevant information for learning the Markov chain parameters from this data.

Inference. For trees, the CGM distribution can be written in closed form (Sundberg, 1975):

$$p(\mathbf{n}) = M! \prod_{i\in V} \prod_{x_i} \left(\frac{n_i(x_i)!}{\mu_i(x_i)^{n_i(x_i)}} \right)^{\nu_i - 1} \cdot \prod_{(i,j)\in E} \prod_{x_i, x_j} \frac{\mu_{ij}(x_i, x_j)^{n_{ij}(x_i, x_j)}}{n_{ij}(x_i, x_j)!} \tag{1}$$

where $\mu_i(x_i) = \Pr(X_i = x_i)$ and $\mu_{ij}(x_i, x_j) = \Pr(X_i = x_i, X_j = x_j)$ are the marginal probabilities of the graphical model (prior to any observations), and ν_i is the degree of vertex i. The distribution is only supported on the set of consistent node and edge tables $\mathcal{L}_M = \{\mathbf{n} : n_i(x_i) = \sum_{x_j} n_{ij}(x_i, x_j), \forall i, x_i, j \in N(i), \text{ and } \sum_{x_i} n_i(x_i) = M, \forall i\}$ that are nonnegative-integer valued. The MAP inference problem for CGMs is to maximize $p(\mathbf{n} \mid \mathbf{y})$ over this feasible set. By relaxing variables to be real-valued, taking the negative log of the objective, and using Stirling's approximation, Sheldon et al. (2013) arrived at the following convex relaxation of the MAP problem:

$$\min_{\mathbf{n}\in\mathcal{L}_M} f(\mathbf{n}) = -\sum_{(i,j)\in E} \sum_{x_i, x_j} n_{ij}(x_i, x_j) \log \psi_{ij}(x_i, x_j) - H_B(\mathbf{n}) - \sum_{i\in V} \sum_{x_i} \ell_{i,x_i}(n_i(x_i)), \tag{MAP}$$

where $\psi_{ij}(x_i, x_j)$ are new potentials that collect terms that are linear in $\mathbf{n}$ from $\log p(\mathbf{n})$, and $H_B(\mathbf{n})$ is the *Bethe entropy*:

$$H_B(\mathbf{n}) = -\sum_{(i,j)\in E} \sum_{x_i, x_j} n_{ij}(x_i, x_j) \log n_{ij}(x_i, x_j) + \sum_{i\in V} (\nu_i - 1) \sum_{x_i} n_i(x_i) \log n_i(x_i). \tag{2}$$

For trees, the Bethe entropy is concave over $\mathcal{L}_M$ (Heskes, 2006), and thus the overall problem is convex and can be solved by off-the-shelf solvers (Sheldon et al., 2013). This inference approach is extremely accurate and much faster than the previous method of Gibbs sampling, but it is still not efficient enough for large-scale problems.

3 Message Passing Algorithm

Readers familiar with the *Bethe free energy* will recognize the close resemblance between the objective of (MAP) and that function, which is defined as:

$$F_B(\boldsymbol{\tau}) = -\sum_{(i,j)\in E}\sum_{x_i,x_j} \tau_{ij}(x_i,x_j)\log\psi_{ij}(x_i,x_j) - H_B(\boldsymbol{\tau}).$$

The only difference is that the CGM objective has additional convex terms $-\ell_{i,x_i}(n_i(x_i))$ that correspond to the CGM evidence. Yedidia et al. (2000) showed that Pearl's classical belief propagation (BP) algorithm (1988), if it converges, reaches a zero-gradient point of the Lagrangian of the Bethe free energy with respect to the constraint $\boldsymbol{\tau} \in \mathcal{L}_1$ (the set of locally-consistent node and edge marginals that sum to one). BP maintains a set of messages $\{m_{ij}(x_j)\}$ from nodes to their neighbors, which are updated according to the rule:

$$m_{ij}(x_j) \propto \sum_{x_i} \psi_{ij}(x_i,x_j) \prod_{k\in N(i)\setminus j} m_{ki}(x_i), \tag{3}$$

Upon convergence, the node marginals are $\tau_i(x_i) \propto \prod_{k\in N(i)} m_{ki}(x_i)$ and the edge marginals are $\tau_{ij}(x_i,x_j) \propto \psi_{ij}(x_i,x_j)\prod_{k\in N(i)\setminus j} m_{ki}(x_i)\prod_{l\in N(j)\setminus i} m_{lj}(x_j)$ (normalized to sum to one). In practice, if BP converges on a loopy graph, it usually converges to a minimum of the Bethe free energy (Heskes et al., 2003). We remind the reader that, for trees, BP always converges, and that minimizing $F_B(\boldsymbol{\tau})$ over $\mathcal{L}_1$ is the exact variational problem for marginal inference. For graphs with cycles, both the Bethe free energy F_B and the constraint set $\mathcal{L}_1$ are approximations of their exact counterparts (Wainwright and Jordan, 2008).

$$\begin{aligned} n_i(x_i) &\propto \prod_{k\in N(i)} m_{ki}(x_i) \\ \widehat{\psi}_{ij}(x_i,x_j) &= \psi_{ij}(x_i,x_j)\times \\ &\qquad \exp\left\{\frac{1}{\nu_i}\ell'_{i,x_i}(n_i(x_i)) + \frac{1}{\nu_j}\ell'_{j,x_j}(n_j(x_j))\right\} \\ m_{ij}(x_j) &\propto \sum_{x_i}\widehat{\psi}_{ij}(x_i,x_j)\prod_{k\in N(i)\setminus j} m_{ki}(x_i) \end{aligned} \tag{4}$$

Figure 1: CGM message passing

CGM Message Passing. A generalization of the argument of Yedidia et al. (2000) can be used to guess message updates that solve the CGM problem (Figure 1). *The only difference from BP is that the edge potentials are updated in each iteration*: The first line computes the current marginals (normalized to sum to M). The second line updates the edge potentials based on the gradient of the CGM evidence terms evaluated at the current marginals. The final line is the standard BP update.

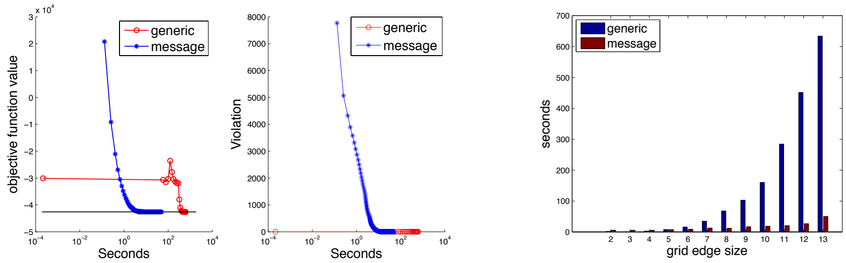
Theorem 1. *Assume G is a tree. If the CGM message updates converge, the resulting vector $\mathbf{n}$ of node and edge marginals is an optimal solution to* (MAP).

Proof sketch. Following Yedidia et al. (2000), we write the Lagrangian of (MAP) and set the gradients with respect to the primal and dual variables to zero to derive first-order optimality conditions. Inspection of these equations reveals the nature of the Lagrange multipliers as messages. Guessing the correspondence between the Lagrange multipliers and messages in the algorithm shows that convergence of the messages is equivalent to satisfaction of the zero-gradient conditions. Because the problem is convex for trees, these conditions are also sufficient for optimality. □

A full proof is deferred to a longer version of the paper. We note that a trick of modifying (MAP) by duplicating the node marginal variables and adding constraints to enforce their meaning greatly facilitates the derivation and reveals the need for the gradient terms in Equation (4). Note that, unlike standard BP, convergence is not guaranteed even for trees. In practice, we found that message damping (Heskes et al., 2003) was needed, and sufficient damping always led to convergence in our experiments.

4 Evaluation

We evaluated our message passing algorithm on synthetic data by comparing the solution quality and running time with those of the MATLAB interior point solver for the problem (MAP). Following the setup of Sheldon et al. (2013), synthetic data was generated from a chain-structured CGM to simulate wind-dependent migration of a population of M birds from the bottom-left to the top-right corner of an $\ell \times \ell$ grid. The variables X_t of the individual model are the grid locations of individual birds at times $t = 1, \ldots, T$, and have cardinality $L = \ell^2$. The transition probabilities between grid cells were determined by a log-linear model with four parameters that control the effect of features such as direction, distance, and wind on the transition probability. The parameters were selected manually to generate realistic migration trajectories. We generate node and edge contingency tables from this process and then generated noisy observations from the Poisson model $y \sim \text{Pois}(\alpha n + \alpha_o)$ where the intensity rate and background rate are set to be 1 and 0 respectively. We generate synthetic data for grids of different sizes to test the scaling behavior of the algorithms.



(a) (b)

Figure 2: Convergence behavior and running time comparison.

Figure 2(a) shows the convergence behaviors of both algorithms for the 13×13 grid. The left plot shows the MAP objective value as a function of time, and the right plot shows the constraint violation. Both algorithms find optimal feasible solutions, but message passing does so much faster, especially in terms objective value. Figure 2(b) compares the running time of the algorithms on different grid sizes. The message passing algorithm clearly scales much better to larger problems.

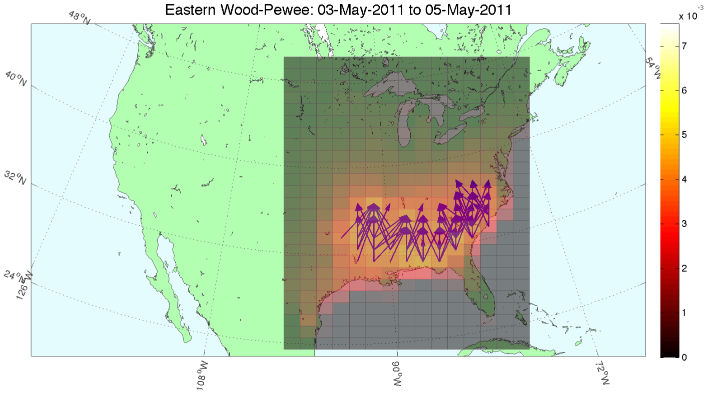


Figure 3: Reconstruction of migration routes.

The increased speed of message passing allows us to apply CGMs back to realistic instances of the motivating bird migration application. We divided the eastern US into a 25×15 grid, and used the STEM species distribution model (Fink et al., 2010) to estimate $y_t(x_t)$, the fraction of the total population of Eastern Wood-pewees in grid cell x_t during week t, for all x_t and for 22 weeks during spring migration. To model uncertainty in STEM estimates, we assume they follow a Gaussian distribution centered around the true value: $y_t(x_t) \sim \mathcal{N}(n_t(x_t), \sigma^2)$, with all values normalized to a hypothetical population of 1000 birds. We used a fixed discretized Gaussian transition model similar to that of Sheldon et al. (2008). Figure 3 shows transition counts $n_t(x_t, x_{t+1})$ recovered from the MAP inference problem for week 9. Arrows indicate transitions made by more the 2% of the population. This is a simple idealized model for bird migration, but it shows empirically that message passing can be used in large-scale problems for learning and inference for CGMs.

Acknowledgment. This material is based upon work supported by the National Science Foundation under Grant No. 1125228.

References

D. Fink, W. Hochachka, B. Zuckerberg, D. Winkler, B. Shaby, M. Munson, G. Hooker, M. Riedewald, D. Sheldon, and S. Kelling. Spatiotemporal exploratory models for broad-scale survey data. *Ecological Applications*, 20(8):2131–2147, 2010.

T. Heskes. Convexity arguments for efficient minimization of the Bethe and Kikuchi free energies. *Journal of Artificial Intelligence Research*, 26(1):153–190, 2006.

T. Heskes et al. Stable fixed points of loopy belief propagation are minima of the bethe free energy. *Advances in neural information processing systems*, 15:359–366, 2003.

J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 1988.

D. Sheldon. *Manipulation of PageRank and Collective Hidden Markov Models*. PhD thesis, Cornell University, 2009.

D. Sheldon and T. G. Dietterich. Collective graphical models. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

D. Sheldon, M. A. S. Elmohamed, and D. Kozen. Collective inference on Markov models for modeling bird migration. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

D. Sheldon, T. Sun, A. Kumar, and T. G. Dietterich. Approximate inference in collective graphical models. In *International Conference on Machine Learning (ICML)*, 2013.

B. L. Sullivan, C. L. Wood, M. J. Iliff, R. E. Bonney, D. Fink, and S. Kelling. eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, 142(10):2282 – 2292, 2009.

R. Sundberg. Some results about decomposable (or Markov-type) models for multidimensional contingency tables: distribution of marginals and partitioning of tests. *Scandinavian Journal of Statistics*, 2(2):71–79, 1975.

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

J. S. Yedidia, W. T. Freeman, and Y. Weiss. Generalized belief propagation. In *NIPS*, volume 13, pages 689–695, 2000.