Inferning with High Girth Graphical Models

Uri Heinemann URIHEI@CS.HUJI.AC.IL
The Hebrew University of Jerusalem, Jerusalem, Israel

Amir Globerson GAMIR@CS.HUJI.AC.IL
The Hebrew University of Jerusalem, Jerusalem, Israel

Abstract

Unsupervised learning of graphical models is an important task in many domains. Although maximum likelihood learning is computationally hard, there do exist consistent learning algorithms (e.g., psuedo-likelihood and its variants). However, inference in the learned models is still hard, and thus they are not directly usable. In other words, given a probabilistic query they are not guaranteed to provide an answer that is close to the true one.

In the current paper, we provide a learning algorithm that is guaranteed to provide approximately correct probabilistic inference. We focus on a particular class of models, namely high girth graphs in the correlation decay regime. It is well known that approximate inference (e.g, using loopy BP) in such models yields marginals that are close to the true ones. Motivated by this, we propose an algorithm that always returns models of this type, and hence in the models it returns inference is approximately correct. We derive finite sample results guaranteeing that beyond a certain sample size, the resulting models will answer probabilistic queries with a high level of accuracy.

Results on synthetic data show that the models we learn indeed outperform those obtained by other algorithms, which do not return high girth graphs.

1. Introduction

Graphical models are a highly useful tool for describing multivariate distributions (Koller and Friedman, 2009). Such models assume that the distribution can be written as a product of functions, each defined over a small set of parameters. When these functions are pairwise, the structure of the model can be described via the edges of a graph.

To fully describe a graphical model , one must specify both its structure and parameters. In principle, once these are available, one can answer probabilistic queries such as: "what is the probability that $X_1 = 0$ given that $X_5 = 0, X_6 = 0$". Indeed the power of graphical models lies in their ability to answer any such probabilistic inference query. Unfortunately, for most models of interest these inference queries are computationally intractable (e.g., $\#P$ hard). The common approach to this state of affairs is to either use approximate inference algorithms (e.g., sampling, loopy belief propagation, mean field) or to focus on families where inference is tractable or can be well approximated (e.g., low tree width graph, or high girth graphs as we focus on here).

The situation is further complicated by the fact that in most cases the graph structure and parameters are not known a priori, and need to be learned from data. Ideally, we would like inference in the learned model to agree with inference in the true underlying distribution. In other words, if the true model is $p^*(x)$ we would like the learned model to answer the probabilistic query $\mathbb{P}\left[X_1 = 0 | X_5 = 0, X_6 = 0\right]$ with a number that is close to $p^*(x_1 = 0 | x_5 = 0, x_6 = 0)$.[1] If inference in the learned model cannot be performed exactly, one needs to consider the interplay between learning and inference, which has recently been referred to as "Inferning".[2]

The above goal of making accurate inferences in the learned model is quite ambitious, and is indeed not achieved by most current learning methods, including ones which have otherwise strong theoretical guarantees. For example the psueodlikelihood method for parameter estimation (Besag, 1975) and various recent methods for struc-

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32. Copyright 2014 by the author(s).

[1] We slightly abuse notation of the PMF by not including indices indicating which conditional distribution is considered.

[2] See: http://inferning.cs.umass.edu/2012.

ture learning (Ravikumar et al., 2010; Jalali et al., 2011) are known to be consistent (i.e., they recover correct structure and parameters as $n \to \infty$). However, inference in the models they learn is generally intractable, and thus there is no reason to expect that it will come close to inference in $p^*(x)$.[3]

One way to make progress towards our goal is to assume that $p^*(x)$ has structure and parameters that facilitate exact or approximately correct inference. In this setting it makes sense to expect that the learned model will inherit these properties (at least asymptotically) and thus yield inference that is close to $p^*(x)$. Perhaps the most elegant illustration of this approach is the celebrated Chow Liu (CL) algorithm (Chow and Liu, 1968) which learns a tree structured graphical model. CL can also be shown to be consistent for both parameter and structure learning (Tan et al., 2011). Since inference in trees is tractable, inference in the learned tree models will approach that of the true model.

However, CL is limited in that it can only use tree structured models, and these are often not sufficient for describing complex dependencies that arise in real data. The next natural step is to look for model classes where inference is not tractable, but approximate inference with approximation bounds is possible. Here we focus on such a class, namely, models with high girth and correlation decay (or HGCD). These models are "tree-like" in the sense of having long cycles, but they exhibit much richer properties. For example, in LDPC codes, high girth structures can be used to approach capacity (Richardson et al., 2001). An attractive property of HGCD is that inference, despite being intractable, can be ϵ approximated (with ϵ depending on the HGCD structure and parameters) using loopy belief propagation (Mezard and Montanari, 2009).

Given the above, there is clear motivation for learning a graphical model that is a HGCD. It is important to emphasize that our goal is not to asymptotically learn such a model, but to return such a model for any given finite sample. This will guarantee that inference in our learned HGCD is ϵ approximate. In turn, it will let us show that if the true distribution $p^*(x)$ is an HGCD, then inference in the learned model will be $O(\epsilon)$ close to $p^*(x)$ given a sufficiently large but finite training sample.

Our contribution is thus as follows: we describe an algorithm that takes as input a sample generated by an HGCD p^* with graph G. The algorithm returns another HGCD with graph $\bar{G}$, which has the following desirable properties as $n \to \infty$: it is guaranteed (whp) to include the edges of G, it is guaranteed to achieve test likelihood that is close to the best possible, and, most importantly, inference in the learned model is guaranteed to be close to inference in p^*.

The algorithm we propose is very simple, and generalizes Chow Liu in a very natural way. It proceeds as follows: for every pair of variables (i, j) set $w_{i,j}$ to be the empirical mutual information between these variables. Now construct G by greedily adding the edges with the highest weight, as long as G has the required girth. After having learned the model structure, we set the parameters in a similar way to CL, while ensuring they are in the HGCD regime.

After describing the algorithm and proving its properties, we illustrate its performance on synthetic data. We show that it indeed performs more accurate predictions than models that are not restricted to high girth.

2. Problem Formulation

Graphical models are used to compactly describe multivariate distributions over p random variables $X_1, \dots, X_p$ (e.g., see Koller and Friedman, 2009). We use $\boldsymbol{x} = (x_1, \dots, x_p)$ to denote an assignment to the p variables. The model $(G, \boldsymbol{\theta})$ is parameterized via a set of edges E, and functions $\theta_{ij}(x_i, x_j)$ for each $ij \in E$. The distribution is given by:

$$p(x; \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} e^{\sum_{ij \in E} \theta_{ij}(x_i, x_j) + \sum_i \theta_i(x_i)} \qquad (2.1)$$

where $Z(\boldsymbol{\theta})$ is the partition function that normalizes the distribution.

To simplify presentation and analysis we focus on the case where $X_i \in \{-1, +1\}$ are binary variables. In this case we obtain the Ising model, which can be parameterized in the following way (here J_{ij}, h_i are scalars):

$$p(x; \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} e^{\sum_{ij \in E} J_{ij} x_i x_j + \sum_i h_i x_i} \qquad (2.2)$$

Thus for the Ising model we have:

$$\theta_i(x_i) = \begin{bmatrix} -h_i \\ h_i \end{bmatrix}, \quad \theta_{ij}(x_i, x_j) = \begin{bmatrix} J_{ij} & -J_{ij} \\ -J_{ij} & J_{ij} \end{bmatrix} \qquad (2.3)$$

Given a sample $\mathcal{D} = \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(n)}$ of n assignments sampled IID from the *true* distribution $p^*(\boldsymbol{x})$, we wish to approximate p^* with a graphical model. The classic approach to learning parametric models from data is by maximizing the likelihood, which in our case would be

$$L(\boldsymbol{\theta}, E) = \frac{1}{n} \sum_{i=1}^{n} \log p(\boldsymbol{x}^{(i)}; \boldsymbol{\theta}) \qquad (2.4)$$

where it is implied that $\boldsymbol{\theta}$ is restricted to the edges E. For general models, this function is hard to optimize and even

[3] We note that many methods learn models that match certain marginals of the empirical data (corresponding to edges of the graph), and are thus guaranteed to accurately recover this small set of marginals. However no guarantees are available for other probabilistic queries. See for example (Wainwright, 2006).

evaluate (since it requires the evaluation of the log partition function). Note that what we would actually like to maximize is the *generalization* likelihood, namely the expected likelihood according to the *true* distribution (e.g., see Dudík et al., 2007):

$$L^*(\boldsymbol{\theta}, E) = \sum_{\boldsymbol{x}} p^*(\boldsymbol{x}) \log p(\boldsymbol{x}; \boldsymbol{\theta}) \tag{2.5}$$

However, since we don't know p^* this cannot be maximized directly. One option is to instead maximize the empirical likelihood Eq. 2.4. However this is often computationally hard, and other methods such as pseudo likelihood may be used. The method presented here also does not maximize the empirical likelihood directly, but is still consistent.

If E is restricted to tree graphs, then the likelihood in Eq. 2.4 can be exactly maximized using the Chow Liu algorithm. CL uses two simple facts. The first is that for any tree E, the optimal values of $\boldsymbol{\theta}$ are given by the so called canonical parameters:

$$\theta_i^c(x_i) = \log \mu_i(x_i) \quad , \quad \theta_i^c(x_i, x_j) = \log \frac{\mu_{ij}(x_i, x_j)}{\mu_i(x_i)\mu_j(x_j)} \tag{2.6}$$

where μ_i, μ_{ij} are the singleton and pairwise marginals calculated from the data. We will denote the set of all canonical parameters by $\boldsymbol{\theta}^c$ or $\boldsymbol{\theta}^c(\mu)$ to highlight their dependence on the marginals μ. Second, for these parameters the likelihood turns out to be:

$$L(\boldsymbol{\theta}^c, E) = \sum_{ij \in E} I_{\mathcal{D}}(X_i; X_j) + \sum_i H_{\mathcal{D}}(X_i) \tag{2.7}$$

where $I_{\mathcal{D}}(X_i; X_j)$ is the empirical mutual information between the two variables, and $H_{\mathcal{D}}(X_i)$ is the empirical entropy. The final observation is that the E that maximizes Eq. 2.7 is the maximum weight spanning tree where edge weights are given by:

$$w_{ij} = I_{\mathcal{D}}(X_i; X_j) \tag{2.8}$$

As mentioned earlier our focus is on high-girth graphs. Recall that a girth of the graph is the length of its shortest cycle. Our goal will be to learn an E that has girth at least g, for some fixed g (we refer to this as a girth bounded graph, implicitly meaning that it is a lower bound). Note that this significantly complicates likelihood maximization, since even if the likelihood can be approximated as a sum of edge weights, finding the maximum weight bounded girth graph is NP hard (Kortsarz et al., 2008). Thus, we proceed using a different approach, as explained next.

3. The Extended Chow Liu Algorithm

We first present our algorithm for learning bounded girth models. Our Extended Chow Liu algorithm (ECL) first learns a structure E and then assigns parameters $\boldsymbol{\theta}$, as explained next.

The structure learning part is a simple extension of CL to graphs with bounded girth. The first stage is to calculate the weights w_{ij} as in Eq. 2.8. Next, the learned E is constructed by greedily adding the edges with largest w_{ij}, as long as the graph has girth lower bounded by g. The procedure is described in Figure 1. Note that if $g = p + 1$, it reduces to standard CL, but otherwise it is different from CL since it does not return a tree. Computationally, ECL requires keeping track of paths in the graph, such that generation of a cycle shorter than g can be efficiently identified. By using standard graph search algorithms, this can be implemented in $O(p^3)$.[4]

Once the structure E has been determined, we turn to setting the parameters $\boldsymbol{\theta}$. If there is no constraint on the interaction strength we set $\boldsymbol{\theta} = \boldsymbol{\theta}^c$ for the edges in $\bar{E}$, as in Eq. 2.6. In other words, the parameters are set exactly as in the Chow Liu algorithm.

In some cases we may want to further constrain the parameters, so that interaction strength is bounded, and we are in the correlation decay regime. This is addressed in Section 6. However, note that most of our theoretical results for learning parameters and structure hold even if we do not return a model with these bounded interactions. The reason is that given enough data our learned parameters will satisfy the bound automatically.

Algorithm 1 Extended Chow Liu

```
Require: $\mathcal{D}, g$
E = $\emptyset$
Calculate $w_{ij} = I_{\mathcal{D}}(X_i; X_j)$ for all $ij$
loop
    Set $S$ to be all $e \notin E$ that satisfy $girth(E \cup e) > g$.
    If $S = \emptyset$ break.
    Find $e \in S$ with max $w_e$ and set $E = E \cup e$.
end loop
Set $\boldsymbol{\theta} = \boldsymbol{\theta}^c$ (see Eq. 2.6).
return $E, \boldsymbol{\theta}$
```

4. Model Assumptions

To provide theoretical guarantees on ECL, we need to make several assumptions about the true underlying distribution. These pertain to both the structure and parameters of the model, as is standard in results of this type.

The following definitions will be needed in what follows.

Definition 1. *For a graph G and nodes i, j, define $B_l(ij)$ to be the set of nodes in G that are at distance less*

[4] When p is large the complexity is $O(d_{max}^{2l})$.

than or equal to l from i or j. Namely: $B_l(ij) = \{k \in V | \min\{d(i,k), d(j,k)\} \leq l\}$ (e.g., see Anandkumar and Valluvan, 2013). Furthermore, define $\partial B_l(ij)$ to be the boundary of $B_l(ij)$, so that: $\partial B_l(ij) = \{k \in V | \min\{d(i,k), d(j,k)\} = l\}$

Definition 2. *Given an $\epsilon > 0$ let $\mathcal{LG}_\epsilon$ define a class of graphical models where there exists an $l \in \mathbb{N}$ such that for all $i \in V$,and all $j \in nei(i)$,*

$$1 - \epsilon < \frac{p(x_i, x_j | x_{\partial B_l(ij)}; \boldsymbol{\theta})}{p(x_i, x_j; \boldsymbol{\theta})} < 1 + \epsilon \qquad (4.1)$$

and $B_l(ij)$ is a tree.

Thus $\mathcal{LG}_\epsilon$ are models where the effect of neighbors beyond l is small, and the l neighborhood is tree structured. In other words, the models in $\mathcal{LG}_\epsilon$ are locally tree like and exhibit correlation decay. Similar assumption appear in the literature on belief propagation (Mezard and Montanari, 2009).[5]

Our assumptions on the learned model are as follows (we will soon see that they imply the model is in $\mathcal{LG}_\epsilon$). Given an $0 < \epsilon < 0.01$ we say that $(G, \boldsymbol{\theta})$ satisfies the A_ϵ assumptions if:

A1: For all $i \in V$ the parameter h_i satisfies:

$$|h_i| \leq h_{max} \leq \frac{1}{2} \ln(\epsilon^{-\frac{1}{2}} - 1) - 1 \qquad (4.2)$$

A2: Denote $l = \lfloor \frac{g-1}{2} \rfloor$, where g is the girth of G and $\eta = \frac{1}{1+e^{2h_{max}+2}}$. Then $\forall ij \in E$, the parameters J_{ij} satisfy:

$$|J_{ij}| \leq J_{\max} \equiv \tanh^{-1}\left(\frac{1}{2d_{max}} \epsilon^{2/l} \eta^2\right) \qquad (4.3)$$

where d_{max} is the maximum degree of the graph.

A3: All edges $ij \in E$ satisfy $I_{\boldsymbol{\theta}}(X_i; X_j) \geq 13\epsilon$.

The assumptions above are similar to those used elsewhere (e.g., see Ravikumar et al., 2010; Anandkumar and Valluvan, 2013). They basically rule out interactions that are too strong or too weak, as well as singleton marginals that are too peaked. The scale for the interaction strength is determined with respect to the girth of the graph. As the girth increases the J_{ij} become less constrained.

The key property of the models that satisfy the above is that inference can be performed with accuracy ϵ using the belief propagation algorithm. This result has appeared in several variants in the past (Mezard and Montanari, 2009) and is given below.

Lemma 1. *Assume the model $(G, \boldsymbol{\theta})$ satisfies A_ϵ. Then running belief propagation on the model will converge to marginals $\tau_{ij}(x_i, x_j)$ such that for all $ij \in E$:*

$$|p(x_i, x_j; \boldsymbol{\theta}) - \tau_{ij}(x_i, x_j)| \leq \epsilon^2 \qquad (4.4)$$

See proof in Section 1 of the supplementary file.

The above restrictions on $(G, \boldsymbol{\theta})$ imply that the model is in the correlation decay regime, as the following lemma states.

Lemma 2. *If $(G, \boldsymbol{\theta})$ satisfies the A_ϵ assumptions above, then the model is in $\mathcal{LG}_\epsilon$.*

See proof in Section 2 of the supplementary file.

5. Theoretical Analysis

The goal of this section is to show that ECL results in a model that is close to optimal in terms of the generalization likelihood Eq. 2.5. In other words, it learns a model whose likelihood is close to the likelihood obtained by the true model. Our Theorem 3 will state that for all ϵ, if the true model satisfies A_ϵ, then we can get ϵ close to the optimal likelihood. We perform the analysis in both the $n \to \infty$ sample size limit and for finite samples (e.g., see Tan et al., 2011, for related results).

We proceed in two steps. First, we show that the graph learned by the model will contain the true graph (see Section 5.1). Next, assuming that the graph is correct, we show that the learned parameters will results in a likelihood that is at worse ϵ suboptimal.

5.1. Structure Consistency

We first address the question of finding the correct graph structure of the Ising model using ECL. We begin with the case of $n \to \infty$ IID samples, and show the structure can be exactly recovered, up to errors on low information edges.[6] In the infinite data setting, ECL will receive as input the correct mutual information values $I_{\boldsymbol{\theta}}(X_i; X_j)$. The question then becomes whether this greedy procedure will recover the correct structure given the correct model information. The result is provided in the following theorem.

Theorem 1. *Let $(G^*, \boldsymbol{\theta}^*)$ be an Ising model satisfying assumptions A_ϵ. Denote the model graph by $G = (V^*, E^*)$. Then running ECL with mutual informations $I_{\boldsymbol{\theta}^*}(X_i; X_j)$ calculated from $p(x; \boldsymbol{\theta}^*)$ and the true girth of G, will result in a set of edges E such that:*

- *If $ij \in E^*$ then $ij \in E$. Namely, E contains all edges in E^*.*

[5] Note that this is a different definition from (Anandkumar and Valluvan, 2013), since we require a multiplicative factor and they require an additive one.

[6] If our goal is structure learning, we can discard these edges by thresholding their information. However this is not our focus here.

- *If $ij \in E \setminus E^*$ then $I_{\boldsymbol{\theta}^*}(X_i; X_j) \leq 13\epsilon$. Namely, E contains no redundant edges except possibly those with mutual information less than 13ϵ.*

The proof is based on the following two lemmas. The first states that edges from i to outside $B_l(ij)$ indeed have small information. See proof in Section 3.1 of the supplementary file.

Lemma 3. *Assume $(G, \boldsymbol{\theta})$ satisfies A_ϵ. Then for all $ij \in E$ and $k \notin B_l(ij)$ it holds that $I_{\boldsymbol{\theta}}(X_i; X_k) \leq \epsilon^2$.*

The following lemma facilitates the greedy ECL algorithm. It states that in our model the information between non-edges is always less than the information of all edges on the path between them. See proof in Section 3.2 of the supplementary file.

Lemma 4. *Assume $(G, \boldsymbol{\theta})$ satisfies A_ϵ. Let $ij \notin E$ be two nodes whose distance in G is $q < \lfloor \frac{g-1}{2} \rfloor$. Let $P^{ij} = \{x_i = x_{p_1}, \ldots, x_{p_q} = x_j\}$ be a shortest path in the graph between i and j. Then:*

$$I_{\boldsymbol{\theta}}(X_i; X_j) + 3\epsilon < I_{\boldsymbol{\theta}}(X_{p_s}; X_{p_{s+1}}) \;\forall 1 \leq s \leq q-1 \quad (5.1)$$

We can now intuitively see why Theorem 1 holds. The ECL algorithm greedily adds edges as long as they do not form cycles that violate the girth constraint. According to Lemma 3 above, edges outside the tree neighborhood of a node have low information and thus will not be added (since edges in the tree neighborhoods will have larger information). Within the tree neighborhood Lemma 4 states that edges that “shortcut” paths in the tree will have lower information than edges in the path, and thus these “non-tree” edges will also not be chosen. A more formal proof is given in the supplementary file.

5.2. Structure Estimation from Finite Samples

Next we consider the case where the mutual informations used by ECL are calculated from a finite sample. Intuitively, as more data becomes available the information estimates should improve and we should be able to find the correct structure with high probability. The following theorem quantifies the number of samples. See proof in Section 3.4 of the supplementary file.

Theorem 2. *Assume an IID sample of size n is generated from a model $(G^*, \boldsymbol{\theta}^*)$ satisfying A_ϵ. For any $\delta > 0$ let n satisfy $n > N_0$ where:*

$$N_0 = \frac{C_0 \log p \log \frac{1}{\delta}}{\epsilon^3} \quad (5.2)$$

and C_0 is a constant. Then with probability greater than $1 - \delta$ ECL will recover E^* as in Theorem 1.

The proof uses Chernoff bounds to verify that the empirical mutual informations are sufficiently close to their true values, and Lemma 4 which bounds the difference between different informations, and thus determines the level of resolution required in estimating these.

5.3. Likelihood Optimality

We have thus far proven that the learned graph will contain the true graph with high probability. However, our goal is to obtain a complete model where inference can be performed. Thus, we need to also ask how well the parameters can be learned . Theorem 3 below states that the likelihood of the learned model will not be too far from the optimal one. See proof in Section 4.1 of the supplementary file.

Theorem 3. *Assume a sample is generated IID from a model $(G^*, \boldsymbol{\theta}^*)$ satisfying A_ϵ where $\epsilon < \frac{1}{|E^*|-p}$. Then as $n \to \infty$ ECL will return $(G, \boldsymbol{\theta})$ such that*

$$L^*(\theta^*, E^*) - L^*(\theta, E) < 2\epsilon \quad (5.3)$$

where L^* is the generalization likelihood (see Eq. 2.5).

Note that the above is equivalent to the Kullback Leibler divergence between p^* and the learned model being small. Namely:

$$D_{KL}[p(x; \theta^*) | p(x; \theta)] \leq 2\epsilon \quad (5.4)$$

Informally, Theorem 3 follows from the following facts:

- The canonical parameters used by ECL maximize the Bethe likelihood (e.g., see Heinemann and Globerson, 2011).[7]
- The canonical parameters also satisfy A_ϵ, and are thus in $\mathcal{LG}_\epsilon$.
- For models in $\mathcal{LG}_\epsilon$ the true likelihood and Bethe likelihood are ϵ close.

5.4. Likelihood Optimality from Finite Samples

As in the structure learning case, we can obtain finite sample bounds that guarantee the result in Theorem 3 with high probability. The following theorem provides such a bound.

Theorem 4. *Assume an IID sample of size n is generated from a model $(G^*, \boldsymbol{\theta}^*)$ satisfying A_ϵ where $\epsilon < \frac{1}{|E^*|-p}$. For any $\alpha > 0, \delta > 0$, let n satisfy $n > N_1$ where:*

$$N_1 = \frac{C_1 |E^*| \log \frac{1}{\delta}}{\epsilon^3 \alpha^2} \,. \quad (5.5)$$

and C_1 is a constant. Then ECL will return a model that satisfies the following with probability greater than $1 - \delta$:

$$L^*(\theta^*, E^*) - L^*(\theta, E) < 3\epsilon + \alpha \,, \quad (5.6)$$

where L^* is the generalization likelihood (see Eq. 2.5).

[7] Generally canonical parameters are not guaranteed to maximize the Bethe likelihood, but here they do because of the assumptions A_ϵ.

We note that there are other methods for consistently learning the parameters of the model. For example, pseudo likelihood can learn the correct parameters given that $n \to \infty$. Pseudo likelihood in fact has better $n \to \infty$ behavior, since it is not ϵ suboptimal like our method (e.g., see Bradley and Guestrin, 2012). However, several factors make our canonical parameter approach more attractive in practice. First, the sample complexity of psuedo-likelihood is inferior to ours since the former needs a reliable estimate of the Markov blanket of each node. Indeed, our experiments show that ECL requires less data to learn. Second, the canonical parameters are evaluated simply and in closed form from the empirical marginals, and do not require an optimization procedure like pseudo-likelihood. Finally, in the model learned by pseudo-likelihood there are no guarantees on inference quality since they are not HGCD.

6. ECL with correlation decay parameters

Our original motivation was that high girth models in the correlation decay regime can yield high accuracy inference. However, until now we only made sure to return bounded girth models, but with no guarantee on correlation decay. Here we show how ECL can be further modified so that it is guaranteed to return such models.

Recall that a model is guaranteed to be in $\mathcal{LG}_\epsilon$ if its parameters satisfy $|J_{ij}| < \zeta$, for an appropriately defined ζ (as in assumption A_ϵ). We would thus like to return such a set of parameters. It turns out we can learn such a model and preserve all other properties of our algorithm.

The parameter estimates in ECL (i.e., the canonical parameters in Eq. 2.6) are inspired by the parameters that maximize the likelihood for tree structured models. It thus makes sense to ask what are the parameters that maximize likelihood for tree models under the constraint $|J_{ij}| < \zeta$. It turns out that there is a closed form expression for these, as described next. See proof in Section 5 of the supplementary file.

Theorem 5. *Assume we are given a set of empirical marginals $\bar{\mu}$, and a tree structured graph $G = (V, E)$. Then the parameters $\boldsymbol{\theta}^*$ that maximize the likelihood under constraints $|J_{ij}| \le \zeta$ are as follows. Define:*

$$\begin{aligned} \bar{\alpha}_{ij} &= \bar{\mu}_{ij}(1,1) - \bar{\mu}_i(1)\bar{\mu}_j(1) \\ \lambda^+_{ij} &= \max\{0, \bar{\alpha}_{ij} - \phi(\zeta, \bar{\mu}_i, \bar{\mu}_j)\} \\ \lambda^-_{ij} &= \max\{0, -\bar{\alpha}_{ij} + \phi(-\zeta, \bar{\mu}_i, \bar{\mu}_j)\} \end{aligned}$$

where ϕ is the function defined in Section 5 of the supplementary file. Use these to define a new set of marginals $\tilde{\bar{\mu}}(x_i, x_j)$ given as follows:[8]

$$\begin{cases} \bar{\mu}_{ij}(x_i,x_j) & |\alpha_{ij}| \le \phi(\zeta, \bar{\mu}_i\bar{\mu}_j) \\ \bar{\mu}_{ij}(x_i,x_j) - \lambda^+_{ij} & x_i = x_j \text{ and } \alpha_{ij} > \phi(\zeta, \bar{\mu}_i, \bar{\mu}_j) \\ \bar{\mu}_{ij}(x_i,x_j) + \lambda^+_{ij} & x_i \neq x_j \text{ and } \alpha_{ij} > \phi(\zeta, \bar{\mu}_i, \bar{\mu}_j) \\ \bar{\mu}_{ij}(x_i,x_j) + \lambda^-_{ij} & x_i = x_j \text{ and } \alpha_{ij} < \phi(-\zeta, \bar{\mu}_i, \bar{\mu}_j) \\ \bar{\mu}_{ij}(x_i,x_j) - \lambda^-_{ij} & x_i \neq x_j \text{ and } \alpha_{ij} < \phi(-\zeta, \bar{\mu}_i, \bar{\mu}_j) \end{cases} \tag{6.1}$$

Then $\boldsymbol{\theta}^*$ are the canonical parameters calculated from marginals $\tilde{\bar{\mu}}$, namely $\boldsymbol{\theta}^* = \boldsymbol{\theta}^c(\tilde{\bar{\mu}})$.

Thus, if we want ECL to return parameters bounded by ζ will simply use the above parameters instead of $\theta^c(\bar{\mu})$. Assuming the true parameters satisfy this upper bound (as in assumption A_ϵ) then it's easy to see that all the results proven so far will still hold (i.e., structure and parameter consistency).

The advantage of the above scheme is that it will result in models in $\mathcal{LG}_\epsilon$ and thus guarantee ϵ approximate inference using loopy BP. On the other hand, in cases where the model does not satisfy assumption A_ϵ it may be too strict a constraint to limit ζ as it will limit the correlation that can be modeled using the learned model.

We denote the ECL algorithm with the above parameter bounding scheme by ECL_B.

7. Accuracy on Probabilistic Queries

As mentioned in Section 1, the end goal of learning is to answer probabilistic queries. Of course we would like these answers to be close to the answer of the same query when performed on the true distribution $p^*(x)$. Next we show that this can indeed be achieved by our method, as the following theorem states.

Theorem 6. *Assume a sample of size n is generated IID from a model $(G^*, \boldsymbol{\theta}^*)$ satisfying A_ϵ where $\epsilon < \frac{1}{|E^*|-p}$. A parameter $\boldsymbol{\theta}$ is learned using the ECL$_B$ algorithm with this sample. To answer the probabilistic query $P[x_i|\boldsymbol{x_o}]$ we run loopy BP on $p(x;\boldsymbol{\theta})$ and denote the resulting probability by $\tau_i(x_i|\boldsymbol{x_o};\boldsymbol{\theta})$. Then for every $\alpha > 0$ if $n > N_1$ (see Eq. 5.5) the following holds with probability greater than $1-\delta$:*

$$\boldsymbol{E}_{p(\boldsymbol{x_o};\boldsymbol{\theta}^*)}[D_{KL}[p(x_i|\boldsymbol{x_o};\boldsymbol{\theta}^*)|\tau_i(x_i|\boldsymbol{x_o};\boldsymbol{\theta})]] \le 4\epsilon + 2\alpha \tag{7.1}$$

The above states that the inferred probabilities are close in D_{KL} to the true ones, when averaging over the evidence $\boldsymbol{x_o}$ using the true distribution. This averaging is sensible since if some $\boldsymbol{x_o}$ is unlikely to appear which do not want to penalize for inference errors in this case. Note that we have provided the result for inferring the probability of a

[8] Note that these will be consistent and feasible by construction.

single variable x_i given a set of variables $\boldsymbol{x_o}$ of arbitrary size. It's possible to give a similar result for inference on larger sets of variables. We now turn to the proof, which simply follows from the previous results.

Proof: By Theorem 3 we have that $D_{KL}[p(\boldsymbol{x};\boldsymbol{\theta}^*)|p(\boldsymbol{x};\boldsymbol{\theta})] \leq 2\epsilon$. Since conditioning and marginalization only reduce the D_{KL} we have that $\boldsymbol{E}_{p(\boldsymbol{x_o};\boldsymbol{\theta}^*)}[D_{KL}[p(x_i|\boldsymbol{x_o};\boldsymbol{\theta}^*)|p(x_i|\boldsymbol{x_o};\boldsymbol{\theta})]] \leq 2\epsilon$. Performing inference using evidence $\boldsymbol{x_o}$ is equivalent to generating a new, smaller, model from $p(\boldsymbol{x};\boldsymbol{\theta})$ with the evidence "folded" in. We will need to prove that this new smaller model is also in $\mathcal{LG}_\epsilon$, so that inference in it is well approximated using LBP. Looking at the proof of Lemma 2 we can see that a model is in $\mathcal{LG}_\epsilon$ even if we remove the restriction on the singletons. Hence the new model is indeed in $\mathcal{LG}_\epsilon$ since the original model satisfies A_ϵ. By a similar argument to the proof of Lemma 1 we have that $1-\epsilon \leq \frac{p(x_i)}{\tau_i(x_i)} \leq 1+\epsilon$. Using the inequality $-log(1-\epsilon) \leq 2\epsilon$ for $\epsilon < 0.5$ we have the result. ▮

8. Experiments

Recall that our goal was to learn models that allow more accurate inference (although not exact) on arbitrary test queries. In this section we compare ECL to other learning algorithms with respect to this performance measure. The following baselines are evaluated:

- The structure learning algorithm of (Ravikumar et al., 2010). This uses L_1 regularization and logistic regression to learn the structure of a graphical model. After learning the model structure, we use psuedo-likelihood to learn its parameters. For inference we use loopy BP. We also experimented with using the canonical parameters $\boldsymbol{\theta}^c$ on the learned structure, but the psuedo-likelihood parameters performed better (possibly since the learned structured is not constrained to be high girth). We denote this method by $L1$.

- The structure learning algorithm of Anandkumar and Valluvan (2013). This procedure constructs tree neighborhoods for each variable and takes their union. It is guaranteed to be consistent under similar conditions to our approach, namely high girth graphs with correlation decay. Similar to the $L1$ baseline, we use psuedolikelihood to learn the model parameters, and then use loopy BP for inference. Here again using canonical parameters resulted in worse performance. The method requires a parameter r which depends on the true parameters. Since we run on synthetic models, we provide it with the true value of r. We call this the TU (for tree union) baseline.

We compare the above two methods to our ECL approach. The ECL algorithm requires a single parameter, namely a lower bound g on the girth of the graph. As with the TU method, since we are using synthetic data, we provided the correct value of this parameter.[9] For ECL, we used the structure learning algorithm Figure 1, and used loopy belief propagation for inference.

Our focus is inference quality in the learned model, rather than learning its model structure. Thus, we evaluate on the accuracy with which the models perform inference. Specifically, we generate 100 random queries as follows: take 5 variables, set their values randomly, and calculate the posterior singleton marginals of the remaining variables. Since we use relatively small models, we can compare the results to the correct posterior marginals.

All the models considered have $p = 20$ variables, so as to allow exact inference for comparisons. The underlying graphs were constrained to have a girth of $g = 8$. This was done by starting with a random tree structure and then adding random edges until the girth was achieved. The field parameters h_i were drawn from a uniform distribution on $[-0.1, 0.1]$. The scale of the interaction parameters J_{ij} varied, as described next. Note that generally for these models, the assumption A_ϵ does not hold. Thus, we are in fact operating in the harder "agnostic" setting.

In what follows we discuss results with respect to the effect of sample size and model parameters.

8.1. Effect of Sample Size

All the baselines considered are known to converge to the correct structure given enough data and assumptions on the true model (the $L1$ method works under more general conditions than the first two). The psuedolikelihood parameters we use for $L1$ and TU are also consistent. The parameters used by ECL are $O(\epsilon)$ optimal under our A_ϵ assumptions. Here we study the effect of sample size on the quality of the inferred marginals. The parameters J_{ij} were drawn from a uniform distribution on $[-1.1, 1.1]$.

Figure 1 shows accuracy as a function of sample size. It can be seen that ECL outperforms the other baselines consistently for all sample sizes considered. Furthermore, as expected, performance improves with sample size.

8.2. Effect of Interaction Parameter

To test the effect of model parameters, we vary the range of the Ising interaction parameters J_{ij}. Specifically, their val-

[9] However, we have noticed in many cases that one can search for g by trying different values, and finding ones for which the learned model reproduces the marginals of the data. In other words, it achieves moment matching. See (Heinemann and Globerson, 2011).

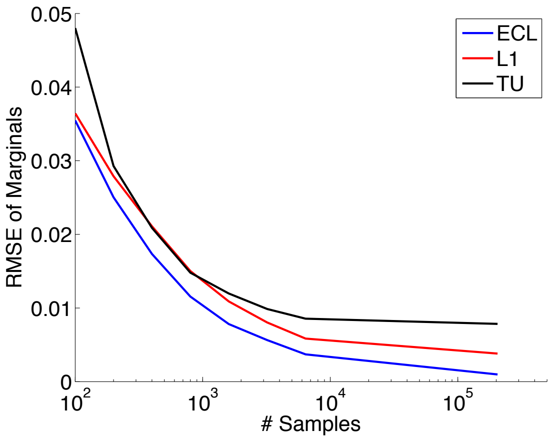

Figure 1. Inference accuracy of learned model as a function of sample size.

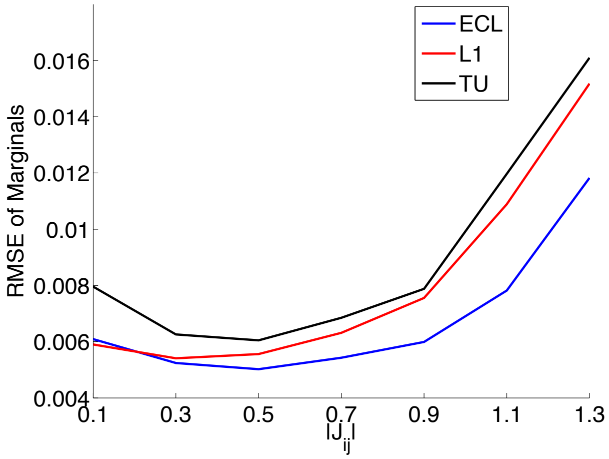

Figure 2. Inference accuracy of learned model as a function of the scale of the interaction parameter.

ues are drawn uniformly from $[-c, c]$ and different values of c are explored. The graph structure is as in the previous sections, and the number of samples is always $n = 3200$.

Figure 2 shows that ECL consistently outperforms the baselines across all parameter values. For small interaction values, the model is close to independent and the baselines all perform similarly. However, as interaction grows, inference in the learned models becomes harder, and thus our high girth models are likely to perform better.

9. Discussion

Our motivation for this work was to learn graphical models that can provide accurate inference at test time. We noted that despite the existence of consistent algorithms for parameter and structure learning, most of these return models that cannot be used for reliable inference.

As an instance of this general approach of learning usable models, we showed how one can learn high girth models with bounded parameters. The advantage is that inference in these models is theoretically guaranteed to provide answers that are close to exact (with closeness being measured by ϵ in the A_ϵ assumptions). We provided consistency results which show that the method both returns high girth models, and is guaranteed to find the correct model structure as well as come close to the optimal likelihood and perform accurate inferences, when the true model is also an HGCD.

Our empirical results demonstrate the utility of our method, which indeed provides better test time inference performance than other baselines with consistency properties.

One limitation of the analysis presented here is that it assumes the true distribution is an HGCD model. However, it seems like if one wants to obtain guarantees on test time accuracy such a restriction must be made. For example, if the true distribution is not constrained, inference in it will be intractable and it is unlikely that any method will be able to approximate it well. Indeed, most methods that provide guarantees on structure learning make assumptions on the true distribution (e.g., see Anandkumar et al., 2012).

The above "inferning" approach can be extended in many ways. Basically, any model class that has theoretical approximation guarantees may be the goal of such an approach. One exciting candidate class are expander graphs, which also exhibit nice theoretical properties for inference (e.g., see Sipser and Spielman, 1996; Burshtein and Miller, 2001). Learning these may turn out to be harder than the high girth case, since it is harder to test for expansion. Another interesting family is fast mixing models (Domke and Liu, 2013). It seems natural to design algorithms which return such models, resulting in guarantees similar to what we presented here.

Acknowledgments: This work was supported by the ISF Centers of Excellence grant 1789/11, by the Israel Ministry of Science and Technology - Knowledge Center in Machine Learning and Articial Intelligence, and by the Intel Collaborative Research Institute for Computational Intelligence (ICRI- CI).

References

A. Anandkumar and R. Valluvan. Learning loopy graphical models with latent variables: Efficient methods and guarantees. *The Annals of Statistics*, 41(2):401–435, 2013.

A. Anandkumar, V. Y. Tan, F. Huang, and A. S. Willsky. High-dimensional structure estimation in Ising models: Local separation criterion. *The Annals of Statistics*, 40 (3):1346–1375, 2012.

J. Besag. The analysis of non-lattice data. *The Statistician*, 24:179–195, 1975.

J. K. Bradley and C. Guestrin. Sample complexity of composite likelihood. In N. D. Lawrence and M. Girolami, editors, *AISTATS*, volume 22 of *JMLR Proceedings*, pages 136–160. JMLR.org, 2012.

D. Burshtein and G. Miller. Expander graph arguments for message-passing algorithms. *Information Theory, IEEE Transactions on*, 47(2):782–790, 2001.

C. I. Chow and C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14:462–467, 1968.

J. Domke and X. Liu. Projecting ising model parameters for fast mixing. In *Advances in Neural Information Processing Systems 26*, pages 665–673, 2013.

M. Dudík, S. Phillips, and R. Schapire. Maximum entropy density estimation with generalized regularization and an application to species distribution modeling. *J. Mach. Learn. Res.*, 8:1217–1260, 2007.

U. Heinemann and A. Globerson. What cannot be learned with Bethe approximations. In *Proceedings of the 27th Conference Annual Conference on Uncertainty in Artificial Intelligence*, pages 319–326. AUAI Press, 2011.

A. Jalali, C. C. Johnson, and P. K. Ravikumar. On learning discrete graphical models using greedy methods. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 1935–1943. 2011.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

G. Kortsarz, M. Langberg, and Z. Nutov. Approximating maximum subgraphs without short cycles. In *Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques*, pages 118–131. Springer, 2008.

M. Mezard and A. Montanari. *Information, physics, and computation*. OUP Oxford, 2009.

P. Ravikumar, M. J. Wainwright, and J. D. Lafferty. High-dimensional Ising model selection using L_1-regularized logistic regression. *Ann. Statist.*, 38(3):1287–1319, 2010.

T. Richardson, M. Shokrollahi, and R. Urbanke. Design of capacity-approaching irregular low-density parity-check codes. *Information Theory, IEEE Transactions on*, 47 (2):619–637, 2001.

M. Sipser and D. A. Spielman. Expander codes. *Information Theory, IEEE Transactions on*, 42(6):1710–1722, 1996.

V. Y. Tan, A. Anandkumar, L. Tong, and A. S. Willsky. A large-deviation analysis of the maximum-likelihood learning of markov tree structures. *Information Theory, IEEE Transactions on*, 57(3):1714–1735, 2011.

M. Wainwright. Estimating the ”wrong” graphical model: Benefits in the computation-limited setting. *Journal of Machine Learning Research*, 7:1829–1859, 2006.