Latent Graphical Model Selection: Efficient Methods for Locally Tree-like Graphs

Animashree Anandkumar
UC Irvine
a.anandkumar@uci.edu

Ragupathyraj Valluvan
UC Irvine
rvalluva@uci.edu

Abstract

Graphical model selection refers to the problem of estimating the unknown graph structure given observations at the nodes in the model. We consider a challenging instance of this problem when some of the nodes are latent or hidden. We characterize conditions for tractable graph estimation and develop efficient methods with provable guarantees. We consider the class of Ising models Markov on locally tree-like graphs, which are in the regime of correlation decay. We propose an efficient method for graph estimation, and establish its structural consistency when the number of samples n scales as $n = \Omega(\theta_{\min}^{-\delta\eta(\eta+1)-2} \log p)$, where $\theta_{\min}$ is the minimum edge potential, δ is the depth (i.e., distance from a hidden node to the nearest observed nodes), and η is a parameter which depends on the minimum and maximum node and edge potentials in the Ising model. The proposed method is practical to implement and provides flexibility to control the number of latent variables and the cycle lengths in the output graph. We also present necessary conditions for graph estimation by any method and show that our method nearly matches the lower bound on sample requirements.

Keywords: Graphical model selection, latent variables, quartet methods, locally tree-like graphs.

1 Introduction

It is widely recognized that the process of fitting observed data to a statistical model needs to incorporate latent or hidden factors, which are not directly observed. Learning latent variable models involves mainly two tasks: discovering structural relationships among the observed and hidden variables, and estimating the strength of such relationships. One of the simplest models is the *latent class model* (LCM), which incorporates a single hidden variable and the observed variables are conditionally independent given the hidden variable. Latent tree models extend this model class to incorporate many hidden variables in a hierarchical fashion. Latent trees have been effective in modeling data in a variety of domains, such as phylogenetics [1].

Another important property regarding latent trees is their computational tractability: upon learning the latent tree model, inference can be carried out efficiently through *belief propagation*. There has been extensive work on learning latent trees, including some of the recent works, e.g. [2–4], where it is demonstrated that latent trees can be learnt efficiently in high dimensions. In other words, the number of samples required for consistent learning is much smaller than the number of variables at hand.

However, despite all the above advantages, latent trees may not be suitable in all scenarios and the assumption of an underlying tree structure may be too restrictive. For instance, consider the example of topic-word models, where topics (which are hidden) are discovered using information about word co-occurrences. In this case, a latent tree model does not accurately represent the hierarchy of topics and words, since there are many common words across different topics. In this paper, we relax the latent tree assumption to incorporate cycles in the latent graphical model, and at the same time, we retain many advantages of latent tree models, including tractable learning and inference.

Relaxing the tree constraint leads to many challenges: in general, learning these models is NP-hard, even when there are no latent variables, and developing tractable methods for such models is itself an area of active research, e.g. [5–7]. In this paper, we consider structure estimation in latent graphical models Markov on locally tree-like graphs, meaning that local neighborhoods in the graph do not contain cycles. These extensions of latent tree models are relevant in many settings: for instance, when there is a small overlap among different hierarchies of variables, the resulting graph has mostly long cycles. There are many questions to be addressed: are there parameter regimes where these models can be learnt consistently and efficiently? If so, are there practical learning algorithms? Are learning guarantees for loopy models comparable to those for latent trees? How does learning depend on various graph attributes such as node degrees, girth of the graph, and so on?

Our Approach: We consider learning Ising models with latent variables Markov on locally tree-like graphs. We assume that the model parameters are in the regime of correlation decay. In this regime, there are no long-range correlations, and the local statistics converge to a tree limit. The implication of correlation decay is immediately clear: we can employ the available latent tree methods to learn "local" subgraphs consistently, as long as they do not contain any cycles. However, a non-trivial challenge remains: how does one merge these estimated local subgraphs (i.e., latent trees) to obtain an overall graph estimate? Specifically, merging involves matching latent nodes across different latent tree estimates, and it is not clear if this can be performed in an efficient manner.

We employ a different philosophy for building locally tree-like graphs with latent variables. We decouple the process of introducing cycles and latent variables in the output model. We initialize a loopy graph consisting of only the observed variables, and then iteratively add latent variables to local neighborhoods of the graph. We establish correctness of our method under a set of natural conditions. We establish that our method is structurally consistent when the number of samples n scales as $n = \Omega(\theta_{\min}^{-\delta\eta(\eta+1)-2} \log p)$, where p is the number of observed variables, $\theta_{\min}$ is the minimum edge potential, δ is the depth (i.e., graph distance from a hidden node to the nearest observed nodes), and η is a parameter which depends on the minimum and maximum node and edge potentials of the Ising model ($\eta = 1$ for homogeneous models). The sample requirement for our method is comparable to the requirement for many popular latent tree methods, e.g. [2–4]. Moreover, note that when there are no hidden variables ($\delta = 1$), the sample complexity of our method is strengthened to $n = \Omega(\theta_{\min}^{-2} \log p)$, which matches with the sample complexity of existing algorithms for learning fully-observed Ising models [5–7]. Thus, we present an efficient method which bridges structure estimation in latent trees with estimation in fully observed loopy graphical models. Finally, we present necessary conditions for graph estimation by any method and show that our method nearly matches the lower bound.

Our proposed method has a number of attractive features for practical implementation: the method is amenable to parallelization which makes it efficient on large datasets. The method provides flexibility to control the length of cycles and the number of latent variables introduced in the output model. The method can incorporate penalty scores such as the Bayesian information criterion (BIC) [8] to tradeoff model complexity and fidelity. Moreover, by controlling the cycle lengths in the output model, we can obtain models with good inference accuracy under simple algorithms such as loopy belief propagation (LBP). Preliminary experiments on the newsgroup dataset suggests that the method can discover intuitive relationships efficiently, and also compares well with the popular latent Dirichlet allocation (LDA) [9] in terms of topic coherence and perplexity.

Related Work: Learning latent trees has been studied extensively, mainly in the context of phylogenetics. Efficient algorithms with provable guarantees are available (e.g. [2–4]). Our proposed method for learning loopy models is inspired by the efficient latent tree learning algorithm of [4]. Works on high-dimensional graphical model selection are more recent. The approaches can be mainly classified into two groups: non-convex local approaches [5, 6, 10] and those based on convex optimization [7, 11, 12]. There is a general agreement that the success of these methods is related to the presence of correlation decay in the model [13]. This work makes the connection explicit: it relates the extent of correlation decay (i.e., the convergence rate to the tree limit) with the learning efficiency for latent models on large girth graphs. An analogous study of the effect of correlation decay for learning fully observed models is presented in [5]. This paper is the first work to provide provable guarantees for learning discrete graphical models on loopy graphs with latent variables (which can also be easily extended to Gaussian models, see Remark following Theorem 1). The work in [12] considers learning latent Gaussian graphical models using a convex relaxation method, by exploiting a sparse-low rank decomposition of the Gaussian precision matrix. However, the method cannot be easily extended to discrete models. Moreover, the "incoherence" conditions required for the success of convex methods are hard to interpret and verify in general. In contrast, our conditions for success are transparent and based on the presence of correlation decay in the model.

2 System Model

2.1 Ising Models

A *graphical model* is a family of multivariate distributions Markov in accordance to a fixed undirected graph [14]. Each node in the graph $i \in W$ is associated to a random variable X_i taking value in a set $\mathcal{X}$. The set of edges E captures the set of conditional independence relations among the random variables. We say that a set of random variables $\mathbf{X}_W := \{X_i, i \in W\}$ with probability mass function (pmf) P is Markov on the graph G if

$$P(x_i | x_{\mathcal{N}(i)}) = P(x_i | x_{W \setminus i}) \tag{1}$$

holds for all nodes $i \in W$, where $\mathcal{N}(i)$ are the neighbors of node i in graph G.

The Hammersley-Clifford theorem [14] states that under the positivity condition, given by $P(\mathbf{x}_W) > 0$, for all $\mathbf{x}_W \in \mathcal{X}^{|W|}$, a distribution P satisfies the Markov property according to a graph G iff. it factorizes according to the cliques of G. A special case of graphical models is the class of Ising models, where each node consists of a binary variable over $\{-1, +1\}$ and there are only pairwise interactions in the model. In this case, the joint distribution factorizes as

$$P(\mathbf{x}_W) = \exp\left(\sum_{e \in E} \theta_{i,j} x_i x_j + \sum_{i \in V} \phi_i x_i - A(\boldsymbol{\theta})\right), \tag{2}$$

where $\boldsymbol{\theta} := \{\theta_{i,j}\}$ and $\boldsymbol{\phi} := \{\phi_i\}$ are known as edge and the node potentials, and $A(\boldsymbol{\theta})$ is known as the *log-partition function*, which serves to normalize the probability distribution.

We consider latent graphical models in which a subset of nodes is latent or hidden. Let $H \subset W$ denote the hidden nodes and $V \subset W$ denote the observed nodes. Our goal is to discover the presence of hidden variables $\mathbf{X}_H$ and learn the unknown graph structure $G(W)$, given n i.i.d. samples from observed variables $\mathbf{X}_V$. Let $p := |V|$ denote the number of observed nodes and $m := |W|$ denote the total number of nodes.

2.2 Tractable Models for Learning

In general, structure estimation of graphical models is NP-hard. We now characterize a tractable class of models for which we can provide guarantees on graph estimation.

Girth-Constrained Graph Families: We consider the family of graphs with a bound on the *girth*, which is the length of the shortest cycle in the graph. Let $\mathcal{G}_{\rm Girth}(m; g)$ denote the ensemble of graphs with girth at most g. There are many graph constructions which lead to a bound on girth. For example, the bipartite Ramanujan graph [15] and the random Cayley graphs [16] have bounds on the girth. Theoretical guarantees for our learning algorithm will depend on the girth of the graph. However, our experiments reveal that our method is able to construct models with short cycles as well.

Regime of Correlation Decay: This work establishes tractable learning when the graphical model converges locally to a tree limit. A sufficient condition for the existence of such limits is the regime of *correlation decay*, which refers to the property that there are no long-range correlations in the model [5]. In this regime, the marginal distribution at a node is asymptotically independent of the configuration of a growing boundary. For the class of Ising models in (2), the regime of correlation decay can be explicitly characterized, in terms of the maximum edge potential $\theta_{\max}$ of the model and the maximum node degree $\Delta_{\max}$. Define $\alpha := \Delta_{\max} \tanh \theta_{\max}$. When $\alpha < 1$, the model is in the regime of correlation decay, and we provide learning guarantees in this regime.

3 Method, Guarantees and Necessary Conditions

Background on Learning Latent Trees: We first recap some useful techniques employed for latent tree models and then explore on how they can be extended for learning loopy models. Most latent tree learning methods are *distance based*, meaning they are based on the presence of an *additive tree metric* between any two nodes in the tree model. For Ising model (and more generally, any discrete model), the "information" distance between any two nodes i and j in a tree T is defined as

$$d(i, j; T) := -\log|\det(P_{i,j})|, \tag{3}$$

where $P_{i,j}$ denotes the joint probability distribution between nodes i and j. On a tree model T, it can be established that $\{d(i,j)\}$ is additive along any path in T.

Learning latent trees can thus be reformulated as learning tree structure T given end-to-end (estimated) distances $\mathbf{d} := \{\widehat{d}(i,j) : i,j \in V\}$ between the observed nodes V. Various methods with performance guarantees have been proposed, e.g. [2–4]. They are usually based on local tests such as *quartet* tests, involving groups of four nodes. In [4], the so-called CLGrouping method is proposed, which organically grows the tree structure by adding latent nodes to local neighborhoods. In the initial step, the method constructs the minimum spanning tree $\mathrm{MST}(V;\mathbf{d})$ over the observed nodes V using distances $\mathbf{d}$. This tree can be viewed as an attempt to fit all the distances between the observed nodes without using any hidden variables. The method then iteratively visits local neighborhoods of $\mathrm{MST}(V;\mathbf{d})$ and adds latent nodes by conducting local distance tests. Since a tree structure is maintained in every iteration of the algorithm, we can parsimoniously add hidden variables by selecting neighborhoods which locally maximize scores such as the Bayesian information criterion (BIC) [8]. This method also allows for fast implementation by allowing for parallelization of latent tree reconstruction in different neighborhoods, see [17] for details.

Proposed Algorithm: We now propose a method for learning loopy latent graphical models. As in the case of latent tree methods, our method is also based on estimated information distances

$$\widehat{d}^n(i,j;G) := -\log|\det(\widehat{P}^n_{i,j})|, \quad \forall i,j \in V, \tag{4}$$

where $\widehat{P}^n_{i,j}$ denotes the empirical probability distribution at nodes i and j computed using n i.i.d. samples. The presence of correlation decay in the Ising model implies that $\widehat{d}^n(i,j;G)$ is approximately a tree metric when nodes i and j are "close" on graph G (compared to the girth g of the graph). Thus, intuitively, local neighborhoods of G can be constructed through latent tree methods. However, the challenge is in merging these local estimates together to get a global estimate $\widehat{G}$: the presence of latent nodes in the local estimates makes merging challenging. Moreover, such a merging-based method cannot easily incorporate global penalties for the number of latent variables added in the output model, which is relevant to obtain parsimonious representations on real datasets.

We overcome the above challenges as follows: our proposed method decouples the process of adding cycles and latent nodes to the output model. It initializes a loopy graph $\widehat{G}^0$ and then iteratively adds latent variables to local neighborhoods. Given a parameter $r > 0$, for every node $i \in V$, consider the set of nodes $B_r(i;\widehat{\mathbf{d}}^n) := \{j : \widehat{d}^n(i,j) < r\}$. The initial graph estimate $\widehat{G}^0$ is obtained by taking the union of local minimum spanning trees:

$$\widehat{G}^0 \leftarrow \cup_{i\in V} \mathrm{MST}(B_r(i;\widehat{\mathbf{d}}^n);\widehat{\mathbf{d}}^n). \tag{5}$$

The method then adds latent variables by considering only local neighborhoods in $\widehat{G}^0$ and running a latent tree reconstruction routine. By visiting all the neighborhoods, a graph estimate $\widehat{G}$ is obtained. Implementation details about the algorithm are available in [17].

We subsequently establish that correctness of the proposed method under a set of natural conditions. We require that the parameter r, which determines the set $B_r(i;\mathbf{d})$ for each node i, needs to be chosen as a function of the depth δ (i.e., distance from a hidden node to its closest observed nodes) and girth g of the graph. In practice, the parameter r provides flexibility in tuning the length of cycles added to the graph estimate. When r is large enough, we obtain a latent tree, while for small r, the graph estimate can contain many short cycles (and potentially many components). In experiments, we evaluate the performance of our method for different values of r. For more details, see Section 4.

3.1 Conditions for Recovery

We present a set of natural conditions on the graph structure and model parameters under which our proposed method succeeds in structure estimation.

(A1) **Minimum Degree of Latent Nodes:** We require that all latent nodes have degree at least three, which is a natural assumption for identifiability of hidden variables. Otherwise, the latent nodes can be marginalized to obtain an equivalent representation of the observed statistics.

(A2) **Bounded Potentials:** The edge potentials $\boldsymbol{\theta} := \{\theta_{i,j}\}$ of the Ising model are bounded, and let

$$\theta_{\min} \le |\theta_{i,j}| \le \theta_{\max}, \quad \forall\,(i,j) \in G. \tag{6}$$

Similarly assume bounded node potentials.

(A3) **Correlation Decay:** As described in Section 2.2, we assume correlation decay in the Ising model. We require

$$\alpha := \Delta_{\max} \tanh \theta_{\max} < 1, \quad \frac{\alpha^{g/2}}{\theta_{\min}^{\eta(\eta+1)+2}} = o(1), \tag{7}$$

where $\Delta_{\max}$ is the maximum node degree, g is the girth and $\theta_{\min}, \theta_{\max}$ are the minimum and maximum (absolute) edge potentials in the model.

(A4) **Distance Bounds:** We now define certain quantities which depend on the edge potential bounds. Given an Ising model P with edge potentials $\boldsymbol{\theta} = \{\theta_{i,j}\}$ and node potentials $\boldsymbol{\phi} = \{\phi_i\}$, consider its attractive counterpart $\bar{P}$ with edge potentials $\bar{\boldsymbol{\theta}} := \{|\theta_{i,j}|\}$ and node potentials $\bar{\boldsymbol{\phi}} := \{|\phi_i|\}$. Let $\phi'_{\max} := \max_{i \in V} \operatorname{atanh}(\bar{\mathbb{E}}(X_i))$, where $\bar{\mathbb{E}}$ is the expectation with respect to the distribution $\bar{P}$. Let $P(\mathbf{X}_{1,2}; \{\theta, \phi_1, \phi_2\})$ denote an Ising model on two nodes $\{1, 2\}$ with edge potential θ and node potentials $\{\phi_1, \phi_2\}$. Our learning guarantees depend on $d_{\min}$ and $d_{\max}$ defined below.

$$d_{\min} := -\log|\det P(\mathbf{X}_{1,2}; \{\theta_{\max}, \phi'_{\max}, \phi'_{\max}\})|, \quad d_{\max} := -\log|\det P(\mathbf{X}_{1,2}; \{\theta_{\min}, 0, 0\})|, \quad \eta := \frac{d_{\max}}{d_{\min}}.$$

(A5) **Girth vs. Depth:** The depth δ characterizes how close the latent nodes are to observed nodes on graph G: for each hidden node $h \in H$, find a set of four observed nodes which form the shortest *quartet* with h as one of the middle nodes, and consider the largest graph distance in that quartet. The depth δ is the worst-case distance over all hidden nodes. We require the following tradeoff between the girth g and the depth δ:

$$\frac{g}{4} - \delta\eta\,(\eta + 1) = \omega(1), \tag{8}$$

Further, the parameter r in our algorithm is chosen as

$$r > \delta\,(\eta + 1)\,d_{\max} + \epsilon, \quad \text{for some } \epsilon > 0, \quad \frac{g}{4} d_{\min} - r = \omega(1). \tag{9}$$

$(A1)$ is a natural assumption on the minimum degree of the hidden nodes for identifiability. $(A2)$ assumes bounds on the edge potentials. It is natural that the sample requirement of any graph estimation algorithm depends on the "weakest" edge characterized by the minimum edge potential $\theta_{\min}$. Further, the maximum edge potential $\theta_{\max}$ characterizes the presence/absence of long range correlations in the model, and is made exact in $(A3)$. Intuitively, there is a tradeoff between the maximum degree $\Delta_{\max}$ and the maximum edge potential $\theta_{\max}$ of the model. Moreover, $(A3)$ prescribes that the extent of correlation decay be strong enough (i.e., a small α and a large enough girth g) compared to the weakest edge in the model. Similar conditions have been imposed before for graphical model selection in the regime of correlation decay when there are no hidden variables [5].

$(A4)$ defines certain distance bounds. Intuitively, $d_{\min}$ and $d_{\max}$ are bounds on information distances given by the local tree approximation of the loopy model. Note that $e^{-d_{\max}} = \Omega(\theta_{\min})$ and $e^{-d_{\min}} = O(\theta_{\max})$. $(A5)$ provides the tradeoff between the girth g and the depth δ. Intuitively, the depth needs to be smaller than the girth to avoid encountering cycles during the process of graph reconstruction. Recall that the parameter r in our algorithm determines the neighborhood over which local MSTs are built in the first step. It is chosen such that it is roughly larger than the depth δ in order for all the hidden nodes to be discovered. The upper bound on r ensures that the distortion from an additive metric is not too large. The parameters for latent tree learning routines (such as confidence intervals for quartet tests) are chosen appropriately depending on $d_{\min}$ and $d_{\max}$, see [17] for details.

3.2 Guarantees

We now provide the main result of this paper that the proposed method correctly estimates the graph structure of a loopy latent graphical model in high dimensions. Recall that δ is the depth (distance from a hidden node to its closest observed nodes), $\theta_{\min}$ is the minimum (absolute) edge potential and $\eta = \frac{d_{\max}}{d_{\min}}$ is the ratio of distance bounds.

Theorem 1 (Structural Consistency and Sample Requirements) *Under $(A1)$–$(A5)$, the probability that the proposed method is structurally consistent tends to one, when the number of samples scales as*

$$n = \Omega\left(\theta_{\min}^{-\delta\eta(\eta+1)-2} \log p\right). \tag{10}$$

Thus, for learning Ising models on locally tree-like graphs, the sample complexity is dependent both on the minimum edge potential $\theta_{\min}$ and on the depth δ. Our method is efficient in high dimensions since the sample requirement is only logarithmic in the number of nodes p.

Dependence on Maximum Degree: For the correlation decay to hold $(A3)$, we require $\theta_{\min} \leq \theta_{\max} = \Theta(1/\Delta_{\max})$. This implies that the sample complexity is at least $n = \Omega(\Delta_{\max}^{\delta\eta(\eta+1)+2} \log p)$.

Comparison with Fully Observed Models: In the special case when all the nodes are observed[1] $(\delta = 1)$, we strengthen the results for our method and establish that the sample complexity is $n = \Omega(\theta_{\min}^{-2} \log p)$. This matches the best known sample complexity for learning fully observed Ising models [5, 6].

Comparison with Learning Latent Trees: Our method is an extension of latent tree methods for learning locally tree-like graphs. The sample complexity of our method matches the sample requirements for learning general latent tree models [2–4]. Thus, we establish that learning locally tree-like graphs is akin to learning latent trees in the regime of correlation decay.

Extensions: We strengthen the above results to provide non-asymptotic sample complexity bounds and also consider general discrete models, see [17] for details. The above results can also be easily extended to Gaussian models using the notion of *walk-summability* in place of correlation decay (see [18]) and the negative logarithm of the correlation coefficient as the additive tree metric (see [4]).

Dependence on Fraction of Observed Nodes: In the special case when a fraction ρ of the nodes are uniformly selected as observed nodes, we can provide probabilistic bounds on the depth δ in the resulting latent model, see [17] for details. For $\eta = 1$ (homogeneous models) and regular graphs $\Delta_{\min} = \Delta_{\max} = \Delta$, the sample complexity simplifies to $n = \Omega\left(\Delta^2 \rho^{-2} (\log p)^3\right)$. Thus, we can characterize an explicit dependence on the fraction of observed nodes ρ.

3.3 Necessary Conditions for Graph Estimation

We have so far provided sufficient conditions for recovering locally tree-like graphs in latent Ising models. We now provide necessary conditions on the number of samples required by any algorithm to reconstruct the graph. Let $\widehat{G}_n : (\mathcal{X}^{|V|})^n \to \mathcal{G}_m$ denote any deterministic graph estimator using n i.i.d. samples from the observed node set V and $\mathcal{G}_m$ is the set of all possible graphs on m nodes. We first define the notion of the graph edit distance.

Definition 1 (Edit Distance) *Let $G, \widehat{G}$ be two graphs[2] with adjacency matrices $\mathbf{A}_G, \mathbf{A}_{\widehat{G}}$, and let V be the set of labeled vertices in both the graphs (with identical labels). Then the edit distance between $G, \widehat{G}$ is defined as*

$$\operatorname{dist}(\widehat{G}, G; V) := \min_{\pi} ||\mathbf{A}_{\widehat{G}} - \pi(\mathbf{A}_G)||_1,$$

where π is any permutation on the unlabeled nodes while keeping the labeled nodes fixed.

In other words, the edit distance is the minimum number of entries that are different in $\mathbf{A}_{\widehat{G}}$ and in any permutation of $\mathbf{A}_G$ over the unlabeled nodes. In our context, the labeled nodes correspond to the observed nodes V while the unlabeled nodes correspond to latent nodes H. We now provide necessary conditions for graph reconstruction up to certain edit distance.

Theorem 2 (Necessary Condition for Graph Estimation) *For any deterministic estimator $\widehat{G}_m : 2^{\rho m n} \mapsto \mathcal{G}_m$ based on n i.i.d. samples, where $\rho \in [0, 1]$ is the fraction of observed nodes and m is the total number of nodes of an Ising model Markov on graph $G_m \in \mathcal{G}_{\text{Girth}}(m; g, \Delta_{\min}, \Delta_{\max})$ on m nodes with girth g, minimum degree $\Delta_{\min}$ and maximum degree $\Delta_{\max}$, for all $\epsilon > 0$, we have*

$$\mathbb{P}[\operatorname{dist}(\widehat{G}_m, G_m; V) > \epsilon m] \geq 1 - \frac{2^{n\rho m} m^{(2\epsilon+1)m} 3^{\epsilon m}}{m^{0.5\Delta_{\min} m} (m - g\Delta_{\max}^{g})^{0.5\Delta_{\min} m}}, \tag{11}$$

under any sampling process used to choose the observed nodes.

Proof: The proof is based on counting arguments. See [17] for details. □

[1] In the trivial case, when all the nodes are observed and the graph is locally tree-like, our method reduces to thresholding of information distances at each node, and building local MSTs. The threshold can be chosen as $r = d_{\max} + \epsilon$, for some $\epsilon > 0$.

[2] We consider inexact graph matching where the unlabeled nodes can be unmatched. This is done by adding required number of isolated unlabeled nodes in the other graph, and considering the modified adjacency matrices [19].

Lower Bound on Sample Requirements: The above result states that roughly

$$n = \Omega\left(\Delta_{\min}\rho^{-1}\log p\right) \tag{12}$$

samples are required for structural consistency under any estimation method. Thus, when $\rho = \Theta(1)$ (constant fraction of observed nodes), polylogarithmic number of samples are necessary $(n = \Omega(\mathrm{poly}\log p))$, while when $\rho = \Theta(m^{-\gamma})$ for some $\gamma > 0$ (i.e., a vanishing fraction of observed nodes), polynomial number of samples are necessary for reconstruction $(n = \Omega(\mathrm{poly}(p))$.

Comparison with Sample Complexity of Proposed Method: For Ising models, under uniform sampling of observed nodes, we established that the sample complexity of the proposed method scales as $n = \Omega(\Delta^2\rho^{-2}(\log p)^3)$ for regular graphs with degree Δ. Thus, we nearly match the lower bound on sample complexity in (12).

4 Experiments

We now employ latent graphical models for topic modeling, i.e., modeling the relationships between various words co-occurring in documents. Each hidden variable in the model can be thought of as representing a topic, and topics and words in a document are drawn jointly from the graphical model. For a latent tree graphical model, topics and words are constrained to form a tree, while loopy models relax this assumption. We conduct some preliminary experiments on 20 newsgroup dataset, and compare it with the popular latent Dirichlet allocation (LDA) model [9]. We evaluate performance in terms of perplexity and topic coherence, used frequently in topic modeling. In addition, we also study tradeoff between model complexity and data fitting through the Bayesian information criterion (BIC) [8].

Dataset: We consider 16,242 binary samples of 100 keywords selected from the 20 newsgroup data. Each binary sample indicates the appearance of the given words in each posting. These samples are divided in to two equal groups, training and test sets for learning and testing purposes.

Methods: We consider a regularized variant of the method proposed earlier for latent graphical model selection. Here, in every iteration, the decision to add hidden variables to a local neighborhood is based on the improvement of the overall BIC score. This allows us to tradeoff model complexity and data fitting. In addition, we obtain better generalization by avoiding overfitting. Note that our proposed method only deals with structure estimation and we use expectation maximization (EM) for parameter estimation. We compare the proposed method with the LDA model[3]. This method is implemented in MATLAB. We used the modules for LBP, made available with UGM[4] package. The LDA models are learnt using the lda package[5].

Performance Evaluation: We evaluate performance based on the test perplexity [20] given by

$$\text{Perp-LL} := \exp\left[-\frac{1}{np}\sum_{k=1}^{n}\log P(\mathbf{x}^{\text{test}}(k))\right], \tag{13}$$

where n is the number of test samples and p is the number of observed variables (i.e., words). Thus the perplexity is monotonically decreasing in the test likelihood and a lower perplexity indicates a better generalization performance. On lines of (13), we also define

$$\text{Perp-BIC} := \exp\left[-\frac{1}{np}\,\mathrm{BIC}(\mathbf{x}^{\text{test}})\right], \quad \mathrm{BIC}(\mathbf{x}^{\text{test}}) := \sum_{k=1}^{n}\log P(\mathbf{x}^{\text{test}}(k)) - 0.5(\mathrm{df})\log n, \tag{14}$$

where df is the degrees of freedom in the model. For a graphical model, we set $\mathrm{df}^{\mathrm{GM}} := m + |E|$, where m is the total number of variables (both observed and hidden) and $|E|$ is the number of edges in the model. For the LDA model, we set $\mathrm{df}^{\mathrm{LDA}} := (p(m-p) - 1)$, where p is the number of observed variables (i.e., words) and $m - p$ is the number of hidden variables (i.e., topics). This is because a LDA model is parameterized by a $p \times (m-p)$ topic probability

[3] Typically, LDA models the counts of different words in documents. Here, since we have binary data, we consider a binary LDA model where the observed variables are binary.

[4] These codes can be downloaded from `http://www.di.ens.fr/~mschmidt/Software/UGM.html`

[5] `http://chasen.org/~daiti-m/dist/lda/`

Method	r	Hidden	Edges	PMI	Perp-LL	Perp-BIC
Proposed	7	32	183	0.4313	1.1498	1.1518
Proposed	9	24	129	0.6037	1.1543	1.1560
Proposed	11	26	125	0.4585	1.1555	1.1571
Proposed	13	24	123	0.4289	1.1560	1.1576
LDA	NA	10	NA	0.2921	1.1480	1.1544
LDA	NA	20	NA	0.1919	1.1348	1.1474
LDA	NA	30	NA	0.1653	1.1421	1.1612
LDA	NA	40	NA	0.1470	1.1494	1.1752

Table 1: Comparison of proposed method under different thresholds (r) with LDA under different number of topics (i.e., number of hidden variables) on 20 newsgroup data. For definition of perplexity based on test likelihood and BIC scores, and PMI, see (13), (14), and (15).

matrix and a $(m-p)$-length Dirichlet prior. Thus, the BIC perplexity in (14) is monotonically decreasing in the BIC score, and a lower BIC perplexity indicates better tradeoff between model complexity and data fitting. However, the likelihood and BIC score in (13) and (14) are not tractable for exact evaluation in general graphical models since they involve the partition function. We employ loopy belief propagation (LBP) to evaluate them. Note that it is exact on a tree model and approximate for loopy models. In addition, we also evaluate topic coherence, frequently considered in topic modeling. It is based on the average pointwise mutual information (PMI) score

$$\overline{\mathrm{PMI}} := \frac{1}{45|H|} \sum_{h \in H} \sum_{\substack{i,j \in \mathcal{A}(h) \\ i<j}} \mathrm{PMI}(X_i; X_j), \quad \mathrm{PMI}(X_i; X_j) := \log \frac{P(X_i = 1, X_j = 1)}{P(X_i = 1) P(X_j = 1)}, \tag{15}$$

where the set $\mathcal{A}(h)$ represents the "top-10" words associated with topic $h \in H$. The number of such word pairs for each topic is $\binom{10}{2} = 45$, and is used for normalization. In [21], it is found that the PMI scores are a good measure of human evaluated topic coherence when it is computed using an external corpus. It is also observed that using a related external corpus gives a high PMI. Hence, in our experiments, we choose a corpus containing news articles from the NYT articles bag-of-words dataset [22]. This dataset has a vocabulary of 102660 words from 300,000 separate articles. For LDA models, the top 10 words for each topic are selected based on the topic probability vector. For latent graphical models, we used the criterion of information distance to select the top 10 words for each topic.

Experimental Results on Graph Structure: We employ our method to learn the graph structures under different thresholds $r \in \{7, 9, 11, 13\}$, which controls the length of cycles. At $r = 13$, we obtain a latent tree and for all other values, we obtain loopy models. The the first long cycle appears at $r = 9$. At $r = 7$, we find a combination of short and long cycles. We find that models with cycles are more effective in discovering intuitive relationships. For instance, in the latent tree $(r = 13)$, the link between "computer" and "software" is missing due to the tree constraint, but is discovered when $r \leq 9$. Moreover, we see that common words across different topics tend to connect the local subgraphs, and thus loopy models are better at discovering such relationships. The graph structures from the experiments are available in [17].

Experimental Results on Perplexity and Topic Coherence: In Table 1, we present results under our method and under LDA modeling. For the LDA model, we vary the number of hidden variables (i.e., topics) as $\{10, 20, 30, 40\}$. In contrast, our method is designed to optimize for the number of hidden variables, and does not need this input. We note that our method is competitive in terms of both perplexity and topic coherence.We find that topic coherence (i.e., PMI) for our method is optimal at $r = 9$, where the graph has a single long cycle and a few short cycles.

The above experiments confirm the effectiveness of our approach for discovering hidden topics, and are in line with the theoretical guarantees established earlier in the paper. Our analysis reveals that a large class of loopy graphical models with latent variables can be learnt efficiently.

Acknowledgement

The first author is supported in part by the NSF Award CCF-1219234, by the AFOSR Award FA9550-10-1-0310, by the ARO Award W911NF-12-1-0404, and by the setup funds at UCI. The second author is supported by the ONR award N00014-08-1-1015.

References

[1] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge Univ. Press, 1999.

[2] P. L. Erdös, L. A. Székely, M. A. Steel, and T. J. Warnow. A few logs suffice to build (almost) all trees: Part i. *Random Structures and Algorithms*, 14:153–184, 1999.

[3] E. Mossel. Distorted metrics on trees and phylogenetic forests. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pages 108–116, 2007.

[4] M.J. Choi, V.Y.F. Tan, A. Anandkumar, and A. Willsky. Learning latent tree graphical models. *J. of Machine Learning Research*, 12:1771–1812, May 2011.

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

[6] A. Jalali, C. Johnson, and P. Ravikumar. On learning discrete graphical models using greedy methods. In *Proc. of NIPS*, 2011.

[7] P. Ravikumar, M.J. Wainwright, and J. Lafferty. High-dimensional Ising Model Selection Using l1-Regularized Logistic Regression. *Annals of Statistics*, 2008.

[8] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

[9] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. *J. of Machine Learning Research*, 3:993–1022, 2003.

[10] G. Bresler, E. Mossel, and A. Sly. Reconstruction of Markov random fields from samples: some observations and algorithms. In *Intl. workshop APPROX Approximation, Randomization and Combinatorial Optimization*, pages 343–356. Springer, 2008.

[11] N. Meinshausen and P. Bühlmann. High dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

[12] V. Chandrasekaran, P.A. Parrilo, and A.S. Willsky. Latent Variable Graphical Model Selection via Convex Optimization. *Arxiv preprint*, 2010.

[13] J. Bento and A. Montanari. Which Graphical Models are Difficult to Learn? In *Proc. of Neural Information Processing Systems (NIPS)*, Vancouver, Canada, Dec. 2009.

[14] 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.

[15] F.R.K. Chung. *Spectral graph theory*. Amer Mathematical Society, 1997.

[16] A. Gamburd, S. Hoory, M. Shahshahani, A. Shalev, and B. Virag. On the girth of random cayley graphs. *Random Structures & Algorithms*, 35(1):100–117, 2009.

[17] A. Anandkumar and R. Valluvan. Learning Loopy Graphical Models with Latent Variables: Efficient Methods and Guarantees. *Under revision from Annals of Statistics. Available on ArXiv:1203.3887*, Jan. 2012.

[18] A. Anandkumar, V. Y. F. Tan, F. Huang, and A. S. Willsky. High-Dimensional Gaussian Graphical Model Selection: Walk-Summability and Local Separation Criterion. *Accepted to J. Machine Learning Research, ArXiv 1107.1270*, June 2012.

[19] H. Bunke and G. Allermann. Inexact graph matching for structural pattern recognition. *Pattern Recognition Letters*, 1(4):245–253, 1983.

[20] D. Newman, E.V. Bonilla, and W. Buntine. Improving topic coherence with regularized topic models. In *Proc. of NIPS*, 2011.

[21] David Newman, Sarvnaz Karimi, and Lawrence Cavedon. External evaluation of topic models. In *Proceedings of the 14th Australasian Computing Symposium(ACD2009)*, page 8, Sydney, Australia, December 2009.

[22] A. Frank and A. Asuncion. UCI machine learning repository, 2010.