Learning Markov Structure by Maximum Entropy Relaxation

Jason K. Johnson, Venkat Chandrasekaran and Alan S. Willsky*
Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
Cambridge, MA 02139

Abstract

We propose a new approach for learning a sparse graphical model approximation to a specified multivariate probability distribution (such as the empirical distribution of sample data). The selection of sparse graph structure arises naturally in our approach through solution of a convex optimization problem, which differentiates our method from standard combinatorial approaches. We seek the maximum entropy relaxation (MER) within an exponential family, which maximizes entropy subject to constraints that marginal distributions on small subsets of variables are close to the prescribed marginals in relative entropy. To solve MER, we present a modified primal-dual interior point method that exploits sparsity of the Fisher information matrix in models defined on chordal graphs. This leads to a tractable, scalable approach provided the level of relaxation in MER is sufficient to obtain a thin graph. The merits of our approach are investigated by recovering the structure of some simple graphical models from sample data.

1 Introduction

Graphical models offer a convenient representation for multivariate probability distributions and convey the Markov structure in distributions compactly. In such models, a probability distribution is defined with respect to a graph; the vertices of this graph represent random variables, and the edge structure specifies the conditional independence (Markov) properties among the variables. However, the Markov structure in a set of variables is rarely known in advance. Hence, *learning* such structure given an empirical distribution of a set of variables is an important problem. Also, it is often of interest to identify a simpler, more tractable approximation to a given graphical model. We develop an approach that is well-suited to both problems.

The problem of learning the Markov structure of a probability distribution has been extensively studied from the point of view of solving a combinatorial optimization problem. Given a distribution p^* (for example, an empirical distribution obtained from data samples), one searches over a collection of graphs in order to identify a simple graph that still provides a good approximation to p^* in the sense of information divergence. Essentially, this involves projecting the distribution to each candidate graph (minimizing information divergence) and picking the closest one. Several methods focus on chordal graphs due to the fact the projection onto a chordal graph has a simple solution. For instance, it is tractable to find the best tree [1] or to find an optimal sub-graph of a given thin chordal graph [2]. However, the general problem of finding the best k-width graph is NP-complete for $k > 1$ [3], and so heuristic methods are used [4, 5, 6, 7].

In this paper, we propose a novel approach to solve the graphical model selection problem using a convex program as opposed to a combinatorial approach. Our formulation is motivated by the maximum entropy (ME) principle [8, 9]. The ME principle states that subject to linear constraints on a set of statistics, the entropy-maximizing distribution among *all* distributions lies in the exponential family based on those statistics used to define the constraints. Loosely, this suggests that entropy, when used as a maximizing objective function, implicitly favors Markov models which possess as few conditional dependencies as possible while still satisfying the constraints. Proceeding with this point of view, we propose a maximum entropy relaxation (MER) problem in which linear constraints on marginal moments are replaced by a set of nonlinear, convex constraints that enforce closeness to the marginal distributions of p^* in the sense of infor-

*{jasonj,venkatc,willsky}@mit.edu

mation divergence. Roughly speaking, we expect that when p^* is close to a lower-order family of Markov models defined on some graph, the MER approach will automatically "thin" the model, that is, the relaxed probability distribution will lie in that Markov sub-family.

Several methods have recently appeared [10, 11, 12][1] using ℓ_1-penalized information projections, where an ℓ_1-norm on model parameters is used to favor sparse graphs. It is known that these methods are dual to the maximum-entropy method using ℓ_∞ moment constraints [13], which is similar to our MER approach. However, the MER constraints are expressed in terms of relative entropy, and we consider the information-theoretic approach to be more principled. In particular, the MER distribution is invariant to reparameterization of the exponential family.

To solve the MER problem, we develop a scalable algorithm that exploits sparse computations on chordal graphs. This algorithm actually solves a sequence of MER problems based on subsets of the constraints. At each step of the procedure, we add more active constraints (the ones which have the largest constraint violation) until all the constraints that were omitted are found to be inactive. Each MER sub-problem may be formulated with respect to a chordal graph which supports the current constraint set. We solve these sub-problems using a primal-dual interior point method that exploits sparsity of the Fisher information matrix over chordal graphs. Very importantly, this incremental approach to solution of MER still finds the global MER solution in the complete model, but in a manner which exploits sparsity of the MER solution. We focus our development on two classes of exponential families, namely, the Gaussian and Boltzmann models.

The rest of this paper is organized as follows. In Section 2, we provide a brief background on graphical models, exponential families, information theory and relevant properties of chordal graphs. In Section 3, we formulate the MER problem, discuss its model thinning property, and develop efficient algorithms. Simulation results are presented in Section 4 for both the Gaussian and Boltzmann models. We conclude in Section 5 and discuss possible extensions.

2 Preliminaries

2.1 Graphical Models

A graph (or hypergraph) $(V, \mathcal{G})$ consists of a set of vertices $V = \{1, \dots, n\}$ and associated edges $\mathcal{G} \subset \binom{V}{2}$ (or $\mathcal{G} \subset 2^V$) that link vertices together. Here, $\binom{V}{2}$ represents unordered pairs of vertices (pairwise edges) and 2^V represents arbitrary subsets of vertices (hyper-edges). A graphical model is a collection of random variables $x = (x_v)_{v \in V}$ with probability distribution $p(x)$ that is *Markov* with respect to the graph: Given arbitrary subsets $A, B, S \subset V$ such that any path from a vertex in A to a vertex in B necessarily passes through a vertex in S (in other words, S is a *separator* of A and B), the subset of variables $x_A = (x_v)_{v \in A}$ is conditionally independent of x_B given x_S.

A *clique* $C \in \mathcal{C}(\mathcal{G})$ is a vertex set $C \subset V$ in which each pair of vertices are linked by an edge. For a strictly positive probability distribution p that is Markov with respect to $\mathcal{G}$, the Hammersley-Clifford theorem (HC) [14] states that p can be factored in terms of local functions defined on cliques as

$$p(x) = \frac{1}{Z(\psi)} \prod_{C \in \mathcal{C}(\mathcal{G})} \psi_C(x_C), \tag{1}$$

where $\psi_C(x_C)$ depends only on the variables x_C, and $Z(\psi)$ is the *partition function*, which serves to normalize the probability distribution. Conversely, if p can be factored as in (1), then it is Markov on $\mathcal{G}$.

2.2 Exponential Families

We consider parametric families of probability distributions with support $\mathbb{X}^n$ defined by

$$p_\theta(x) = \exp\{\theta^T \phi(x) - \Phi(\theta)\}, \tag{2}$$

where $\phi : \mathbb{X}^n \to \mathbb{R}^d$ are the *sufficient statistics*, θ are the *exponential parameters*, and $\Phi(\theta) = \log \int \exp(\theta^T \phi(x)) dx$ is the *cumulant generating function* (or the *log-partition function*). The family is defined by the set of all *normalizable* $\theta \in \Theta \subset \mathbb{R}^d$, such that $\Phi(\theta) < \infty$, and is said to be *regular* if Θ contains an open neighborhood of $\mathbb{R}^d$. The statistics ϕ are *minimal* if they, together with $\phi_\emptyset(x) = 1$, are linearly independent on $\mathbb{X}^n$. The *moments* $\eta = \mathbb{E}\{\phi(x)\}$ define another parameterization of the family. The set of moments that are realized by the family is denoted $\mathcal{M}$. For regular families with minimal statistics, the map $\Lambda : \Theta \to \mathcal{M}$ defined by the moment calculation $\Lambda(\theta) \triangleq \mathbb{E}_\theta\{\phi(x)\}$ is bijective. We refer the reader to the references for background on exponential families and information geometry [15, 16, 17].

A family of graphical models is obtained by defining "local" statistics, $\phi_E(x_E)$, such that each statistic is a function of just a subset of variables. Our focus here is on two particular families, namely, the Boltzmann and Gaussian models.

[1] Two of these [11, 12] appeared after we submitted the initial draft of this paper.

Boltzmann Model Boltzmann machines are graphical models defined over hypergraphs $\mathcal{G} \subset 2^V$ where all variables are binary valued $x_v \in \{0,1\}$. The family of all probability distributions with support on $\{0,1\}^n$ can be parameterized as:

$$p(x) \propto \exp \sum_{E \subset V} \theta_E \phi_E(x), \tag{3}$$

with sufficient statistics:

$$\phi_E(x) = \prod_{v \in E} x_v \tag{4}$$

that are indicator functions for the event $\{x_v = 1, \forall v \in E\}$. Thus, the moment parameters are given by the probabilities $\eta_E \triangleq \mathbb{E}\{\phi_E(x)\} = \Pr(\{x_v = 1, \forall v \in E\})$ for each $E \subset V$. The functions Λ and Λ^{-1} may both be evaluated using a Möbius transform described briefly in Appendix A (see [18] for further details). These computations are $\mathcal{O}(n2^n)$ using a recursive algorithm.

Gaussian Model Gaussian graphical models [4, 19] are defined over graphs $\mathcal{G} \subset \binom{V}{2}$. We consider the graphical model learning problem involving zero-mean[2] Gaussian distributions. Such distributions are usually parameterized in terms of the symmetric, positive-definite covariance matrix $P = \mathbb{E}\{xx^T\}$ as

$$p(x) \propto \exp\{-\tfrac{1}{2} x^T P^{-1} x\}. \tag{5}$$

The exponential family representation of this model is $p(x) \propto \exp\{-\frac{1}{2} x^T J x\}$ based on the *information matrix* $J = P^{-1}$. Defining sufficient statistics ϕ as

$$\phi(x) = (x_v^2)_{v \in V} \cup (x_u x_v)_{\{u,v\} \subset \binom{V}{2}} \tag{6}$$

we obtain θ and η parameters that are respectively given by elements of the J and P matrices:

$$\theta = (-\tfrac{1}{2} J_{v,v})_v \cup (-J_{u,v})_{\{u,v\}} \tag{7}$$

$$\eta = (P_{v,v})_v \cup (P_{u,v})_{\{u,v\}}. \tag{8}$$

Converting between the moment and exponential parameters is equivalent to converting between P and J. This can be achieved by matrix inversion, which, in general, is an $O(n^3)$ computation.

Markov and Marginal Sub-Families Appealing to HC, with $\psi_E(x_E) = \exp\{\theta_E \phi_E(x_E)\}$, the sub-family of Markov distributions on a graph $\mathcal{G}$ corresponds to a flat submanifold $\Theta(\mathcal{G}) \subset \Theta$ defined by sparsity constraints $\theta_E = 0$ for all subsets $E \notin \mathcal{C}(\mathcal{G})$. This Markov sub-family is an exponential family based on the statistics $\phi_{\mathcal{G}} \triangleq (\phi_E)_{E \in \mathcal{G}}$ with reduced parameterizations $\theta_{\mathcal{G}} \triangleq (\theta_E)_{E \in \mathcal{G}}$ and $\eta_{\mathcal{G}} \triangleq (\eta_E)_{E \in \mathcal{G}}$.

A key feature of both the Boltzmann and Gaussian models is that the marginal distribution $p_C(x_C)$ on any clique $C \in \mathcal{C}$ can be represented within the family. Let $\mathcal{G}_C \triangleq \{E \in \mathcal{G} : E \subset C\}$. The moments $\eta_{\mathcal{G}_C} = (\eta_E)_{E \in \mathcal{G}_C}$ specify the marginal distribution $p_C(x_C)$ in the exponential family based on $\phi_{\mathcal{G}_C}$.

Entropy, Divergence and Fisher Information [9, 15] The *entropy* of a probability distribution p is defined $h(p) \triangleq -\mathbb{E}_p\{\log p(x)\}$, which is a measure of the inherent uncertainty or randomness of the random variable x with probability distribution p. The *information divergence* (or *relative entropy*) between two distributions p and q is $d(p,q) \triangleq \mathbb{E}_p\{\log \frac{p(x)}{q(x)}\}$ and is a non-negative measure of contrast between probability distributions that is zero if and only $p(x) = q(x)$ (a.e.). In Boltzmann models these are computed by summation. In Gaussian models, letting P and Q denote covariances, one obtains:

$$h(P) = \tfrac{1}{2}(\log\det P + n \log 2\pi e) \tag{9}$$

$$d(P,Q) = \tfrac{1}{2}\{\mathrm{tr}(PQ^{-1} - I) - \log\det PQ^{-1}\} \tag{10}$$

In an exponential family, entropy as a function of η satisfies a duality relation with the cumulant function,

$$h(\eta) = \min_{\theta \in \Theta} \left\{\Phi(\theta) - \eta^T \theta\right\} = \Phi(\Lambda^{-1}(\eta)) - \eta^T \Lambda^{-1}(\eta) \tag{11}$$

where $\theta = \Lambda^{-1}(\eta)$ is the unique minimizer. It follows that $\nabla h(\eta) = -\theta$ and $\nabla^2 h(\eta) = -(\nabla^2 \Phi(\theta))^{-1}$ for this $\theta = \Lambda^{-1}(\eta)$.

The *Fisher information* of the exponential family with respect to the moment parameters η is a symmetric, positive-definite matrix defined by

$$G(\eta) \triangleq \mathbb{E}_\eta \left\{(\nabla_\eta \log p(x;\eta))(\nabla_\eta \log p(x;\eta))^T\right\}, \tag{12}$$

which plays an important role in variational methods due to the fact that $\nabla^2 h(\eta) = -G(\eta)$.

Information divergence, expressed as a function of the moment parameters μ and ν of its respective arguments, is seen to be the *Bregmann distance* induced by entropy, i.e.,

$$d(\mu,\nu) = \{h(\nu) + \nabla h(\nu)^T(\mu - \nu)\} - h(\mu) \tag{13}$$

Computing first and second derivatives with respect to μ we have:

$$\nabla_\mu d(\mu,\nu) = \Lambda^{-1}(\mu) - \Lambda^{-1}(\nu) \tag{14}$$

$$\nabla_\mu^2 d(\mu,\nu) = G(\mu) \tag{15}$$

Calculation of $G(\eta)$ in the complete Boltzmann model is described in Appendix A and [18]. In the complete

[2]The mean vector does not play a critical role in Gaussian model identification. Therefore, we only consider the zero-mean case without loss of generality.

Gaussian model, it is given by $G_{(ij),(kl)} = J_{i,k}J_{j,l} + J_{i,l}J_{j,k}$, $G_{(ij),k} = J_{i,k}J_{j,k}$ and $G_{i,k} = \frac{1}{2}J_{i,k}^2$ with $J = P(\eta)^{-1}$ (see [20] for derivations).

2.3 Computations on Thin Chordal Graphs

A graph is *chordal* if for each of its cycles of length greater than three there exists an edge not contained in that cycle which links two nodes of the cycle. A *junction tree* of a graph $\mathcal{G}$ is a tree defined on the set of *maximal* cliques $\mathcal{C} \subset \mathcal{C}(\mathcal{G})$ with the following property: For all $C_i, C_j \in \mathcal{C}$, each clique along the unique path between C_i and C_j in the tree contains $C_i \cap C_j$. It is known that a graph is chordal if and only if it has a junction tree. Importantly, distributions that are Markov on a chordal graph can be factored as

$$p(x) = \frac{\prod_{C\in\mathcal{C}} p(x_C)}{\prod_{S\in\mathcal{S}} p(x_S),} \tag{16}$$

where $\mathcal{C}$ is the set of maximal cliques and $\mathcal{S}$ is the collection of edge-wise separators $C_i \cap C_j$ defined by the edges $\{C_i, C_j\}$ of any junction tree of the graph. We say that a chordal graph is *thin* if it has small maximal cliques.

Using (16), entropy can be expressed in terms of marginal entropies on the cliques and separators of a junction tree of the graph. Using the property that the clique marginals are contained within the Gaussian and Boltzmann models, we have:

$$h_{\mathcal{G}}(\eta_{\mathcal{G}}) = \sum_{C\in\mathcal{C}} h_C(\eta_{\mathcal{G}_C}) - \sum_{S\in\mathcal{S}} h_S(\eta_{\mathcal{G}_S}). \tag{17}$$

Differentiating both sides with respect to moment parameters using $\nabla h(\eta) = -\Lambda^{-1}(\eta)$ we have

$$\Lambda_{\mathcal{G}}^{-1}(\eta_{\mathcal{G}}) = \sum_{C\in\mathcal{C}} \Lambda_C^{-1}(\eta_{\mathcal{G}_C}) - \sum_{S\in\mathcal{S}} \Lambda_S^{-1}(\eta_{\mathcal{G}_S}) \tag{18}$$

Differentiating again using $D\Lambda^{-1}(\eta) = -\nabla^2 h(\eta) = G(\eta)$ we have

$$G_{\mathcal{G}}(\eta_{\mathcal{G}}) = \sum_{C\in\mathcal{C}} G_C(\eta_{\mathcal{G}_C}) - \sum_{S\in\mathcal{S}} G_S(\eta_{\mathcal{G}_S}). \tag{19}$$

Implicit in (18) and (19) is the padding of the terms on the right with zeroes at appropriate locations. We remark that the calculations h_E, Λ_E^{-1} and G_E all have explicit closed-form expressions on fully connected subsets of nodes $E \in \mathcal{G}$ and are tractable for small subsets, thus enabling efficient computation of (17), (18), and (19) for thin chordal graphs.[3] Also, the sparsity of the Fisher information matrix is important later when we use sparse matrix computations to efficiently perform each step of the primal-dual interior-point method.

[3] Computation of (17-18) is $\mathcal{O}(nw^3)$ and $\mathcal{O}(n2^w)$ respectively in the Gaussian and Boltzmann model, where w is the maximum clique size. Computing the sparse matrix (19) is $\mathcal{O}(nw^4)$ and $\mathcal{O}(n4^w)$.

3 Maximum Entropy Relaxation

We present an exponential family formulation of the MER problem, and discuss its model-thinning property. Using this property, we then develop an incremental method for solving the complete MER problem by solving a sequence of sub-problems defined on tractable sub-graphs of the full constraint set. Each sub-problem is solved using a primal-dual interior point method that exploits tractable calculations on chordal graphs.

3.1 Exponential Family Formulation

Let $\mathcal{F}$ be an exponential family with statistics ϕ and moment parameters $\eta \in \mathcal{M}$. Let p^* be a given probability distribution with corresponding moments $\eta^* \triangleq \mathbb{E}_{p^*}\{\phi(x)\}$. We would like to identify a lower-order Markov approximation of p^* defined on some sparse graph (to be determined) that still provides a good approximation to p^*.

We propose to address this problem by solution of the *Maximum Entropy Relaxation* (MER) problem:

$$\begin{array}{ll} \max_{\eta\in\mathcal{M}} & h(\eta) \\ \text{s.t.} & d_E(\eta, \eta^*) \le \delta_E,\ \forall E \in \mathcal{G} \end{array}$$

where $h(\eta)$ denotes entropy in the complete family, $d_E(\eta, \eta^*) \triangleq d(\eta_{\mathcal{G}_E}, \eta^*_{\mathcal{G}_E})$ is the information divergence in the marginal family on edge E between distributions specified by $\eta_{\mathcal{G}_E}$ and $\eta^*_{\mathcal{G}_E}$, the hypergraph $\mathcal{G}$ serves to specify the constraint set and $\delta_{\mathcal{G}} = (\delta_E)_{E\in\mathcal{G}}$ specify tolerance on marginal divergence. We require that every non-empty subset of an edge in $\mathcal{G}$ is also an edge.

For $\eta^* \in \mathcal{M}$ and $\delta_{\mathcal{G}} > 0$, this problem is strictly feasible. It is a convex optimization problem: the objective $h(\eta)$ is a concave function, each marginal divergence $d_E(\eta, \eta^*)$ is a convex function of η for fixed η^* and the set of realizable moments $\mathcal{M}$ is convex. Using minimal statistics ϕ, it is strictly convex such that the MER solution $\tilde{\eta}$ (when it exists) is unique. Finally, we remark that the solution always exists in the Boltzmann model and, in the Gaussian model, it exists if and only if each node is contained by at least one edge-constraint.

Note also, we have not imposed any Markov constraints on the solution of the MER problem. The hypergraph $\mathcal{G}$ serves simply to summarize the constraint set, and may very well be fully connected. Typically, we will specify $\mathcal{G}$ to be the set of all subsets of V up

to size k; for instance, with $k = 2$ we impose all node and pairwise marginal constraints. We always include a complete set of node constraints.

However, we do have the following result concerning the Markov structure of the MER solution $\tilde{\eta}$. We say that the constraint on edge $E \in \mathcal{G}$ is *active* if $d_E(\tilde{\eta}, \eta^*) = \delta_E$ and is *lax* if $d_E(\tilde{\eta}, \eta^*) < \delta_E$. Let $\tilde{\mathcal{G}} \triangleq \{E \in \mathcal{G} : d_E(\tilde{\eta}, \eta^*) = \delta_E\}$ denote the sub-hypergraph corresponding to active constraints.

Theorem (Model-Thinning) The MER solution is Markov with respect to $\mathcal{G}$. Moreover, it is Markov on $\tilde{\mathcal{G}} \subset \mathcal{G}$ defined by the active constraints.

Proof. The Karush-Kuhn-Tucker (KKT) conditions assert that there exist $\lambda_E \geq 0$ for $E \in \mathcal{G}$ such that

$$\nabla h(\tilde{\eta}) - \sum_{E \in \mathcal{G}} \lambda_E \nabla_{\tilde{\eta}} d_E(\tilde{\eta}, \eta^*) = 0 \tag{20}$$

Also, by complementary slackness, $\lambda_E = 0$ for lax constraints. Hence, using $\nabla h(\tilde{\eta}) = -\Lambda^{-1}(\tilde{\eta})$ and $\nabla_{\tilde{\eta}} d_E(\tilde{\eta}, \eta^*) = \Lambda_E^{-1}(\tilde{\eta}) - \Lambda_E^{-1}(\eta^*)$, we have

$$\Lambda^{-1}(\tilde{\eta}) + \sum_{E \in \tilde{\mathcal{G}}} \lambda_E (\Lambda_E^{-1}(\tilde{\eta}) - \Lambda_E^{-1}(\eta^*)) = 0 \tag{21}$$

where each term $(\Lambda_E^{-1})_S$ is zero if $S \not\subset E$. Then, $\tilde{\theta}_E \triangleq (\Lambda^{-1}(\tilde{\eta}))_E = 0$ for all $E \notin \mathcal{C}(\tilde{\mathcal{G}})$, which implies Markovianity on $\tilde{\mathcal{G}}$ and, hence, on $\mathcal{G} \supset \tilde{\mathcal{G}}$ as well. □

Fundamentally, this is the mechanism which allows us to learn graph structure by solving a convex problem. When constraints are laxly satisfied by the MER solution, the model is automatically "thinned".

3.2 Algorithms for Solving MER

Incremental Approach Note that MER is formulated with respect to the complete exponential family (not assuming any Markov structure in advance). For problems of even moderate size, direct solution of MER in the complete model can become intractable due to the high dimension of the parameter vector η. However, based on the model-thinning property, we conclude that if the solution is actually sparse, it should not be necessary to solve MER in the complete parameterization. Hence, we propose the following algorithm to adaptively identify the subset of active constraints and a corresponding lower-order Markov family containing the MER solution:

1. Set $k = 0$. Start with the disconnected graph $\mathcal{G}^{(0)}$ including only node constraints.

2. Solve the reduced MER sub-problem based on just the constraints included in $\mathcal{G}^{(k)}$ (use the primal-dual method described in the following section).

3. Based on the solution $\tilde{\eta}^{(k)}$, evaluate the constraint violations $g_E = d_E(\tilde{\eta}^{(k)}, \eta^*) - \delta_E$ for all $E \in \mathcal{G} \setminus \mathcal{G}^{(k)}$.

4. If $g_E < 0$ for all $E \in \mathcal{G} \setminus \mathcal{G}^{(k)}$, STOP. Then, $\tilde{\eta} = \tilde{\eta}^{(k)}$ is the MER solution.

5. Otherwise, build $\mathcal{G}^{(k+1)}$ by adding edges to $\mathcal{G}^{(k)}$ corresponding to the K largest, positive constraints violations (if there are less than K such edges, add just the edges corresponding to violated constraints). Set $k \leftarrow k + 1$ and go back to Step 2.

We again emphasize that, provided we continue adding violated constraints until all the remaining constraints are satisfied, the final graph $\mathcal{G}^{(k)}$ contains $\tilde{\mathcal{G}}$ and $\tilde{\eta}^{(k)}$ is therefore the *optimal* solution of the original MER problem in the complete family. Unlike greedy methods used in combinatorial approaches, our method is distinguished by the fact that the solution obtained is optimal with respect to our global MER criterion.

Primal-Dual Method on Thin Chordal Graphs Lastly, we specify an efficient method to solve the MER sub-problem arising in Step 2 of the incremental approach provided the constraint graph $\mathcal{G}^{(k)}$ is sufficiently sparse.[4] We assume that $\mathcal{G}^{(k)}$ has a thin chordal super-graph $\bar{\mathcal{G}}^{(k)}$, and solve the MER problem in the Markov sub-family parameterized by $\eta_{\bar{\mathcal{G}}^{(k)}}$. By embedding the optimization problem in a Markov family based on a chordal graph, we are able to compute entropy and its derivatives in an efficient manner.[5] Moreover, we find that the Fisher information is *sparse* in chordal graphs which provides the basis for efficient implementation of Newton's method, enabling super-linearly convergent methods. Although we embed the problem in a chordal graph, we still only impose constraints over $\mathcal{G}^{(k)}$, and hence, by the model thinning property, this embedding does not alter the MER solution with respect to $\mathcal{G}^{(k)}$.

We use the primal-dual interior point method [21], which solves modified KKT conditions by Newton's method to determine an optimal primal-dual pair $(\eta_{\bar{\mathcal{G}}^{(k)}}, \lambda_{\mathcal{G}^{(k)}})$ where $\lambda_{\mathcal{G}^{(k)}} = (\lambda_E)_{E \in \mathcal{G}^{(k)}}$ are Lagrange multipliers. We refer the reader to [21] for details of the primal-dual algorithm. Here, we only comment on how we exploit sparsity to obtain an efficient approach. The key computational step of the algorithm is the so-

[4] In the conclusion, we propose an extension of our approach to handle cases where the MER solution has an intractable graph.

[5] Note that computing $h(\eta_{\mathcal{G}})$ for a *non-chordal* graph $\mathcal{G}$ is difficult because the moment parameters $\eta_{\mathcal{G}}$ only implicitly specify the probability distribution. In general, this would require solution of a variational problem, either to determine the corresponding $\theta_{\mathcal{G}}$ or to determine the maximum-entropy completion to a chordal graph. Hence, the chordal graph embedding method is a critical element in our approach to maximum entropy modeling.

lution of a system of linear equations $H\Delta\eta_{\bar{\mathcal{G}}^{(k)}} = r$, based on the symmetric, positive-definite matrix

$$H = G + \sum_{E\in\mathcal{G}^{(k)}} \lambda_E \left(G_E + \tfrac{1}{\delta_E - d_E} b_E b_E^T \right) \qquad (22)$$

where G is the Fisher information on the chordal graph, G_E is the marginal Fisher information, d_E is the marginal divergence and $b_E = \Lambda_E^{-1}(\eta) - \Lambda_E^{-1}(\eta^*)$ is the difference in marginal exponential parameters on edge E between η and η^*. The matrix G is sparse, inheriting the sparse structure of the chordal graph through (19). Furthermore, each additional term in (22), corresponding to an edge $E \in \mathcal{G}^k \subset \mathcal{C}(\bar{\mathcal{G}}^{(k)})$, is non-zero only for those indices $u, v \in E$. Hence, the fill-pattern of H is the same as for G and we can compute $\Delta\eta_{\bar{\mathcal{G}}^{(k)}} = H^{-1}r$ efficiently using sparse Cholesky factorization and back-substitution.[6] The primal-dual method also requires computation of $\nabla h(\eta_{\bar{\mathcal{G}}^{(k)}}) = -\Lambda^{-1}(\eta_{\bar{\mathcal{G}}^{(k)}})$, which is given by a tractable computation (18). Although we solve for moment parameters $\tilde{\eta}_{\bar{\mathcal{G}}^{(k)}}$, it is straight-forward to obtain $\tilde{\theta}_{\bar{\mathcal{G}}^{(k)}} \triangleq \Lambda^{-1}(\tilde{\eta}_{\bar{\mathcal{G}}^{(k)}})$, again by (18). Also, due to the model-thinning effect, the converged value of $\tilde{\theta}_{\bar{\mathcal{G}}^{(k)}}$ is zero for all edges not contained in $\mathcal{G}^{(k)}$. In other words, adding fill edges to obtain a chordal super-graph $\bar{\mathcal{G}}^{(k)}$ does not spoil the Markov structure of the MER solution with respect to $\mathcal{G}^{(k)}$.

4 Simulation Results

In this section, we describe the results of simulations that demonstrate the effectiveness of the MER framework in learning the Markov structure of Boltzmann and Gaussian models from sample data. The tolerance parameters used in the MER problem are set in proportion to the number of parameters needed to specify the marginal distribution as follows:

$$\delta_E = \gamma \times \begin{cases} |E| + \binom{|E|}{2}, & \text{Gaussian} \\ 2^{|E|} - 1, & \text{Boltzmann} \end{cases} \qquad (23)$$

Here, $\gamma > 0$ is an overall regularization parameter which controls the trade-off between complexity and accuracy in the resulting MER solution. Our motivation for setting δ proportional to parametric complexity is that, for large sample size, the expectation of $d(\eta, \eta^*)$, where η are the actual moments and η^* are empirical, is approximately equal to the number of parameters divided by the number of samples N_s, which also suggests choosing $\gamma \sim 1/N_s$. In the following examples, we explore the effect of varying γ.

[6] This approach is $\mathcal{O}(nw^6)$ and $\mathcal{O}(n8^w)$ respectively in the Gaussian and Boltzmann model. Iterative methods with $\mathcal{O}(nw^3)$ and $\mathcal{O}(n2^w)$ complexity per iteration are also possible and will be presented in a longer paper.

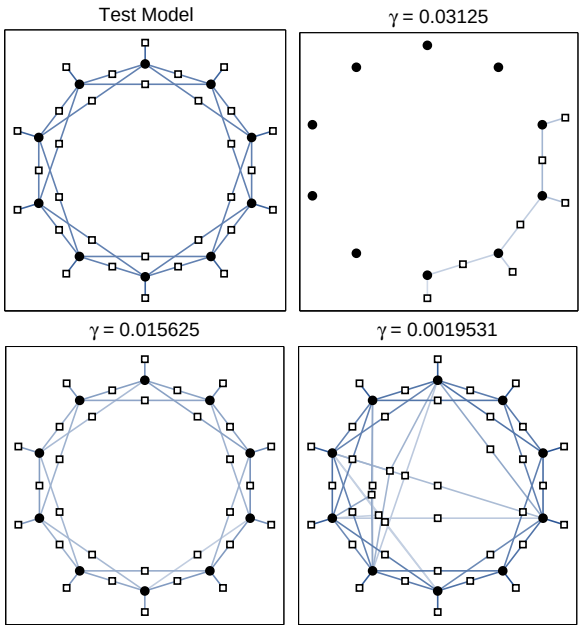

Figure 1: Graphs of the Boltzmann test model and MER of empirical distribution for several values of γ.

In practice, cross-validation methods might prove useful to determine γ that approximately minimizes the generalization error.

Boltzmann model We generated 1000 samples of a 10-node Boltzmann model displayed in Fig. 1. This model includes a pairwise potential $(x_u - \frac{1}{2})(x_v - \frac{1}{2})$ for each pair of vertices that are linked by an edge, which defines a model with unbiased nodes. The empirical moments η^* from these samples were provided as input to MER, where we impose marginal constraints on all singleton, doublet and triplet sets. Fig. 1 shows the MER distribution for several values of γ. Notice the correspondence between the tolerance level and the amount of model-thinning. In this case, for $\gamma = .015625$ we recover the correct graphical structure of the test model.

Gaussian model We describe two sets of experiments for this case. Both simulations were based on 400 samples of test models. In the first experiment, we generate samples from a 16-node cyclic Gaussian model with constant node weights $J_{ii} = -2\theta_i = 1.0$ and edge weights $J_{ij} = -\theta_{ij} = -0.1875$ between nodes that are one or two steps away on the circle. Analogous to the Boltzmann case, Fig. 2 shows the ME relaxation for various values of γ.

The second experiment for the Gaussian involves a 10×10 grid-structured model with edge weight $J_{ij} = -\theta_{ij} = -0.24$ between nearest neighbors in the grid. Again, 400 samples were generated based on this model and the MER problem is solved for a fixed value

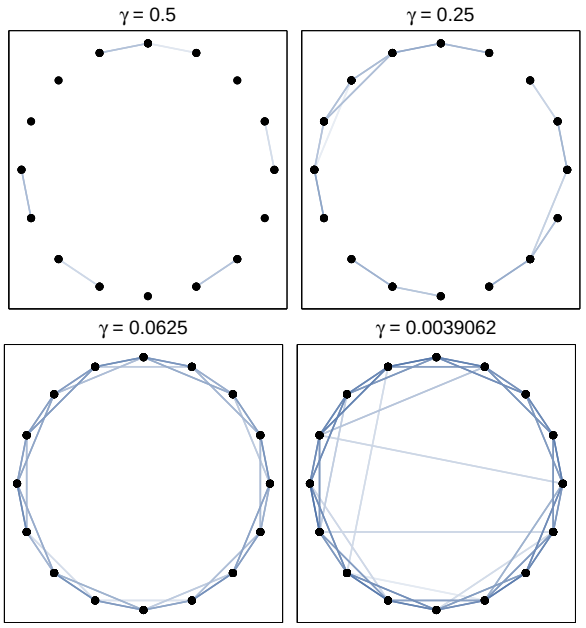

Figure 2: Graphs of the MER solution for various values of γ in the Gaussian model.

of $\gamma = 0.08$. The initial MER problem is solved with 100 node constraints, and at each successive step, the 50 most violated constraints are added. Fig. 3 demonstrates this incremental approach.

In this case, directly solving the MER problem in the complete family (corresponding to the complete graph) would be computationally prohibitive because the resulting Fisher information is a $10,000 \times 10,000$ full matrix. Yet, our incremental approach, using a sequence of thin graphs, solves the MER problem exactly in a few minutes and recovers the underlying graph with very few spurious or missing edges.

5 Conclusion

We have presented a convex optimization approach for learning the graph structure of a collection of random variables from sample data. Our approach differs from previous approaches that addressed this problem primarily from the point of view of solving a combinatorial optimization problem. Our framework is based on the sparsity-enforcing characteristic of entropy (with respect to exponential parameters) that is implicit in the maximum entropy principle. We also exploit sparse, tractable computations of the entropy function and its derivatives on thin chordal graphs in order to solve the MER problem using a scalable primal-dual interior point method. We have demonstrated the effectiveness of our approach with simulation results for the Gaussian and Boltzmann models.

We envision several directions for future research based on the concepts presented in this paper. This paper has been presented primarily from the point of view of learning a model given an empirical distribution. However, our approach can also be used for efficient model-thinning in order to approximate a complex distribution by a lower-order Markov model. This may find applications in recently developed approximate inference methods based on the principle of model thinning [5, 22, 23]. In problems involving a large number of variables, our framework allows for exact solution of the MER problem only when the solution has a tractable graphical structure. If the solution is too complex to compute exactly, it is also of interest to solve the MER problem *approximately* using tractable approximations of the entropy function such as in the Bethe and Kikuchi approximations, or "convexified" versions of these [17].

A Möbius Transform

We briefly discuss the role of the Möbius transform in connecting θ and η parameters in the Boltzmann model. These formulas, derived for the complete Boltzmann model, also apply for marginal computations on cliques and are used to solve MER in thin chordal graphs.

Here, we take $\theta, \eta \in \mathbb{R}^{2^n}$ (adding extra "parameters" $\eta_\emptyset = 1$ and $\theta_\emptyset = -\Phi(\theta)$ associated with the empty set $\emptyset$). These vectors are indexed by subsets $x \subset \{1, \ldots, n\}$ which may be identified with the integers $x \in \{0, \ldots, 2^n - 1\}$ with binary expansion $x = x_n \ldots x_1$ where $x_i = 1$ if i is contained in the subset and $x_i = 0$ otherwise. Also, the vector $p \in \mathbb{R}^{2^n}$ of probabilities of all possible states $x = x_n \ldots x_1$ is indexed similarly.

We define the ω-*transform* of $f \in \mathbb{R}^{2^n}$ by

$$(M_n^\omega f)(x) = \sum_{y \subseteq x} \omega^{|x \setminus y|} f(y) \tag{24}$$

where $|x \setminus y|$ is this number of elements of x that are not contained in y. The *Möbius transform* is given by $M_n \triangleq M_n^1$ and its inverse by M_n^{-1}. One can show by induction that

$$M_n^\omega = \begin{pmatrix} M_{n-1}^\omega & 0 \\ \omega M_{n-1}^\omega & M_{n-1}^\omega \end{pmatrix} \tag{25}$$

where $M_0^\omega = 1$. Hence, we can implement a "fast" ω-transform that requires $n2^n$ computations rather than $\mathcal{O}(2^{2n})$. Given $f = (f_1, f_2) \in \mathbb{R}^{2^n}$ with $f_1, f_2 \in \mathbb{R}^{2^{n-1}}$, we compute the transform of f recursively as $\tilde{f} = (\tilde{f}_1, \omega \tilde{f}_1 + \tilde{f}_2)$. We also use the "upper" Möbius transform M_n^T where the summation is over supersets rather than subsets.

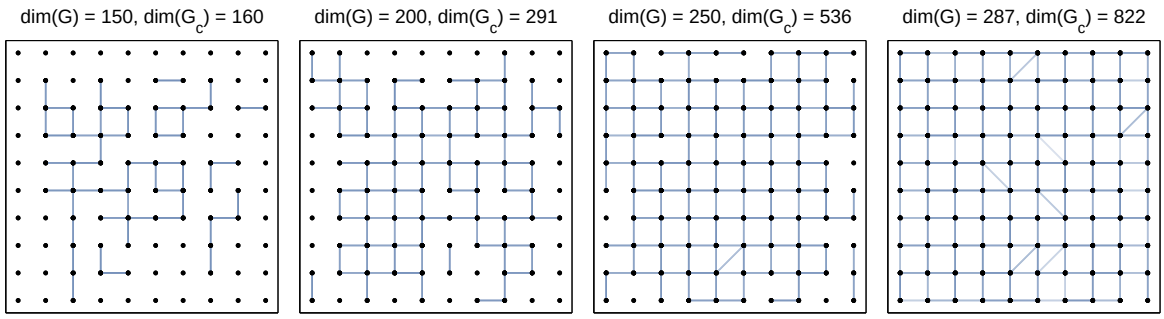

Figure 3: Illustration of the incremental approach for identifying the set of active constraints in the MER problem by solving a sequence of sub-problems defined on sub-graphs of the final solution (far right). Here, $dim(\mathcal{G})$ and $dim(\mathcal{G}_c)$ are the number of nodes plus the number of edges of the constraint graph $\mathcal{G}$ and its chordal super-graph $\mathcal{G}_c$, which respectively determine the dimension of λ and η in each MER sub-problem.

Let $\phi(x) \in \mathbb{R}^{2^n}$ denote the vector of sufficient statistics of the Boltzmann model (with $\phi_\emptyset(x) = 1$) and define $\delta(x)$ to be x-th standard basis vector. We can show the following relations: $\phi(x) = M_n^T \delta(x)$, $\eta = M_n^T p$ and $p = \exp(M_n \theta)$. From these relations, we obtain the following conversion formula:

$$\eta = M_n^T \exp(M_n \theta) \quad (26)$$

$$\theta = M_n^{-1} \log(M_n^{-T} \eta) \quad (27)$$

To obtain the Fisher information matrix, we first compute $G(\eta) \triangleq \frac{\partial \theta}{\partial \eta} = M_n^{-1} \mathrm{Diag}(1/p_\eta) M_n^{-T}$ where $p_\eta = M_n^{-T} \eta$. Because this is using an over-parameterized representation, the Fisher information in the minimal parameterization (without $\eta_\emptyset$) is obtained by deleting the first row and column of G.

References

[1] C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Trans. Information Theory*, 14(3), May 1968.

[2] M. Narasimhan and J. Bilmes. Optimal sub-graphical models. In *NIPS*, 2005.

[3] D. Karger and N. Srebro. Learning Markov networks: maximum bounded tree-width graphs. In *Symposium on Discrete Algorithms*, 2001.

[4] A. Dempster. Covariance selection. *Biometrics*, 28(1), 1972.

[5] U. Kjaerulff. Reduction of computational complexity in Bayesian networks through removal of weak dependencies. In *UAI*, 1994.

[6] F. Bach and M. Jordan. Thin junction trees. In *NIPS*, 2001.

[7] A. Deshpande, M. Garofalakis, and M. Jordan. Efficient stepwise selection in decompasable models. In *UAI*, 2001.

[8] E. Jaynes. Information theory and statistical mechanics. *Physical Review*, 16(4), 1957.

[9] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley & Sons, 1991.

[10] O. Banerjee, L. Ghaoui, A. d'Aspremont, and G. Natsoulis. Convex optimization tehniques for fitting sparse Gaussian graphical models. In *ICML*, 2006.

[11] S. Lee, V. Ganaphthi, and D. Koller. Efficient structure learning of Markov networks using ℓ_1-regularization. In *NIPS*, 2006.

[12] M. Wainwright, P. Ravikumar, and J. Lafferty. Inferring graphical model structure using ℓ_1-regularized pseudo-likelihood. In *NIPS*, 2006.

[13] M. Dudik and R. Schapire. Maximum entropy distribution estimation with generalized regularization. In *COLT*, 2006.

[14] S. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

[15] S. Amari. Information geometry on hierarchy of probability distributions. *IEEE Trans. Information Theory*, 47(5), July 2001.

[16] I. Csiszar. I-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, 3, 1975.

[17] M. Wainwright and M. Jordan. Graphical models, exponential families, and variational inference. Technical Report 649, U.C. Berkeley, 2003.

[18] J. Johnson. On Möbius transforms and Boltzmann machines. unpublished technical note, February 2006.

[19] T. Speed and H. Kiiveri. Gaussian Markov probability distributions over finite graphs. *Annals of Statistics*, 14(1), 1986.

[20] J. Johnson. Fisher information in Gaussian graphical models. unpublished technical note, September 2006.

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

[22] X. Boyen and D. Koller. Tractable inference for complex stochastic processes. In *UAI*, 1998.

[23] J. Johnson. Estimation of GMRFs by recursive cavity modeling. Master's thesis, MIT, March 2003.