Learning Markov Structure by Maximum Entropy Relaxation

Jason K. Johnson, Venkat Chandrasekaran and Alan S. Willsky[†]
Stochastic Systems Group, Laboratory for Information and Decision Systems
Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology
Cambridge, MA 02139, USA
[†]{jasonj,venkatc,willsky}@mit.edu

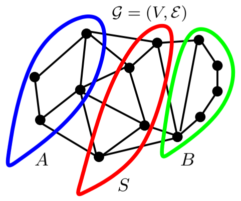
Introduction

- New approach to learning graphical structure using convex optimization rather than combinatorial approach.
- Maximize entropy subject to a set of relaxed marginal constraints, specified in terms of relative entropy on subsets of node variables.
- When constraints are strictly satisfied, the relaxed distribution is "thinned", it is Markov on some sub-graph of the constraint graph.
- We develop a primal-dual interior point method which is tractable on thin graphs.
- To solve the complete problem (e.g., all pairwise constraints) we use an optimistic procedure which incrementally adds most-violated constraints until all constraints are satisfied.

Preliminaries

Graphical Models

A probability model $p(x_1, \dots, x_n)$ is **Markov** on $\mathcal{G}$, with vertices $V = \{1, \dots, n\}$, if separators imply conditional independence:



$$p(x_A, x_B | x_S) = p(x_A|x_S)p(x_B|x_S).$$

Hammersley-Clifford Theorem Markov property is equivalent to existence of factorization over **cliques** (complete sub-graphs):

$$p(x) \propto \prod_{C \in \mathcal{G}} \psi_C(x_C)$$

where $\psi_C(x_C) > 0$ is a function of variables in clique C.

Exponential Families

We consider parametric families of Markov models:

$$p(x;\theta) \propto \exp\left\{\sum_{E \in \mathcal{G}} \theta_E \phi_E(x_E)\right\}, \quad \phi_E(x_E) = \prod_{v \in E} x_v$$

Includes Boltzmann $x \in \{0,1\}^n$ and Gaussian $x \in \mathbb{R}^n$ models. The Gaussian model also includes quadratic statistics $\phi_{v,v}(x) = x_v^2$.

Moment parameters $\eta = \mathbb{E}\{\phi\}$ play central role because exponential family models are **maximum-entropy models** subject to linear moment constraints. The map $\Lambda : \theta \to \eta$ is invertible on the set of realizable moments $\mathcal{M}$ (using a minimal set of features).

Information Theory

Entropy is a measure of uncertainty or randomness in a probability distribution, and is a concave function of η.

$$h(\eta) \triangleq -\mathbb{E}_\eta\{\log p(x;\eta)\}$$

Gradient $\nabla h(\eta) = -\Lambda^{-1}(\eta) = -\theta$, maps η to equivalent θ.

Relative Entropy is a non-negative measure of divergence between probability distributions relative to the curvature of the entropy function (it is the Bregmann distance induced by entropy).

$$d(\mu,\nu) \triangleq \mathbb{E}_\mu\left\{\log\frac{p_\mu(x)}{p_\nu(x)}\right\} = \{h(\nu) + \nabla h(\nu)^T(\mu-\nu)\} - h(\mu) \geq 0$$

Gradient $\nabla_\mu d(\mu,\nu) = \Lambda^{-1}(\mu) - \Lambda^{-1}(\nu)$, difference in θ-coordinates.

Fisher Information Matrix with respect to η is defined:

$$G(\eta) \triangleq \mathbb{E}_\eta\left\{(\nabla_\eta \log p(x;\eta))(\nabla_\eta \log p(x;\eta))^T\right\} = -\nabla^2 h(\eta) \succeq 0$$

Measures curvature of entropy and relative entropy.

Thin Chordal Graphs

A graph is **chordal** if every cycle of four or more nodes is cut by a chord. This holds if and only if there exists a **junction tree**—a tree spanning the maximal cliques of the graph in which the intersection of any two cliques is contained by every clique along the path between them.

In chordal graphs, global entropy calculations reduce to local computations over the cliques and separators (intersections of adjacent cliques) of a junction tree:

$$\begin{aligned} h(\eta) &= \sum_C h_C(\eta_C) - \sum_S h_S(\eta_S) && (1) \\ \Lambda^{-1}(\eta) &= \sum_C \Lambda_C^{-1}(\eta_C) - \sum_S \Lambda_S^{-1}(\eta_S) && (2) \\ G(\eta) &= \sum_C G_C(\eta_C) - \sum_S G_S(\eta_S). && (3) \end{aligned}$$

These computations are tractable on **thin** graphs, where the cliques don't get too large. Also, G is a sparse, thin matrix.

Maximum Entropy Relaxation (MER)

Given moments η^*, we maximize entropy subject to constraints that marginal distributions of small subsets of nodes are close to the marginals of η^* in relative entropy.

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

$\mathcal{G}$ may be a full graph, comprised of all nodes and pairs of nodes; or a hypergraph including all subsets of k or fewer nodes.

Model Thinning The relaxed model is Markov on the sub-graph defined by the active constraints.

Incrementally Add Most-Violated Constraints

We obtain the MER solution by solving a sequence of sub-problems on sub-graphs $\mathcal{G}_k \subset \mathcal{G}$, beginning with the disconnected graph $\mathcal{G}_0$:

1. Solve MER on $\mathcal{G}_k$, parameterized by $(\eta_E)_{E \in \mathcal{G}_k}$ and ignoring constraints not contained in $\mathcal{G}_k$, yielding the relaxation $\tilde{\eta}_k$.
2. Check if $d_E(\tilde{\eta}_k, \eta^*) < \delta_E$ for all $E \in \mathcal{G} \setminus \mathcal{G}_k$. If so then STOP, $\tilde{\eta}_k$ is the MER solution.
3. Otherwise, add the most violated constraints to $\mathcal{G}_k$ to obtain $\mathcal{G}_{k+1}$, set $k \leftarrow k+1$ and go back to Step 1.

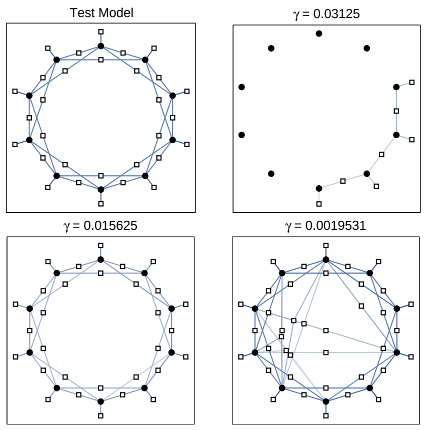
Chordal Embedding and Interior Point Method

- **Chordal Embedding** To solve MER on (non-chordal) $\mathcal{G}_k$, we embed the problem in a chordal graph $\bar{\mathcal{G}}_k \supset \mathcal{G}_k$ and solve MER in the Markov model on $\bar{\mathcal{G}}_k$ parameterized by $(\eta_E)_{E \in \bar{\mathcal{G}}_k}$.
- We use the **primal-dual interior point method**, the state-of-the-art for convex optimization with inequality constraints.
- It applies Newton's method to solve a sequence of modified Karush-Kuhn-Tucker conditions.
- Computing Newton's step reduces to solution of a sparse, linear system based on the sparse, global Fisher information matrix, plus local terms that preserve sparsity.
- Using (1-3) and sparse Cholesky factorization, each step is a tractable $\mathcal{O}(n)$ calculation in bounded tree-width graphs.
- Finally, we recover $\theta_{\bar{\mathcal{G}}_k} = \Lambda^{-1}(\eta_{\bar{\mathcal{G}}_k})$ using (2), which is also sparse (Markov) with respect to $\mathcal{G}_k$.

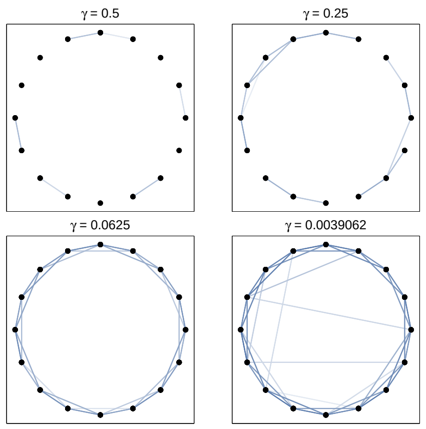
Simulations

We recover the Markov structure of some simple test models by solving MER given the empirical moments η^* from sample data. Each δ_E is proportional to the number of features defined in E, scaled by γ, which controls the level of relaxation.

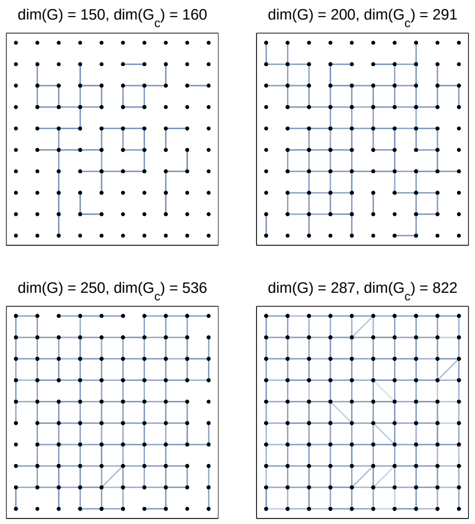
Boltzmann Example



Gaussian Example



Incremental Method for Gaussian Lattice Model



Conclusion

- Convex (non-combinatorial), information-theoretic approach to learning graphical structure in Markov models.
- Tolerance specifications allow trade-off between data-fidelity and sparsity.
- Tractable for learning thin graphs.
- **Future work:** approximate entropy for learning non-thin graphs; hidden variables, inconsistent/incomplete sample data; method for choosing/adapting tolerances; analysis of generalization error and asymptotic performance.

References

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

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

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

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

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

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

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

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

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

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

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