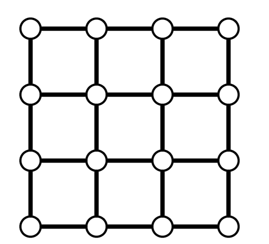
Optimization of Structured Mean Field Objectives

Alexandre Bouchard-Côté* Michael I. Jordan*,†
* Computer Science Division † Department of Statistics
University of California at Berkeley

Inference problem:

$\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^D$
Parameters

$\phi : \mathcal{X} \to \mathbb{R}^D$
Sufficient statistics


Graphical model

$A : \mathbb{R}^D \to \mathbb{R}$
Log-partition

$\mu = \mathbb{E}[\boldsymbol{\phi}(\boldsymbol{X}_\theta)]$
Moments

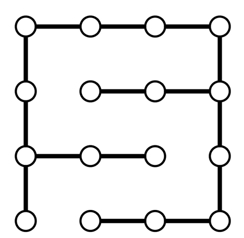
$$\mathbb{P}(\boldsymbol{X}_{\boldsymbol{\theta}} \in B) = \int_B \exp\{\langle \boldsymbol{\phi}(x), \boldsymbol{\theta} \rangle - A(\boldsymbol{\theta})\}\nu(\,\mathrm{d}x)$$

$$A(\boldsymbol{\theta}) = \log \int_{\mathcal{X}} \exp\{\langle \boldsymbol{\phi}(x), \boldsymbol{\theta} \rangle\}\nu(\,\mathrm{d}x)$$

D = # ○ vertex X #{0,1} + # edges X #{00,01,10,11}

Often intractable (e.g. NP-complete for non-planar Ising models)

Overview



Easy special case: dynamic programming can be used when the graph is acyclic

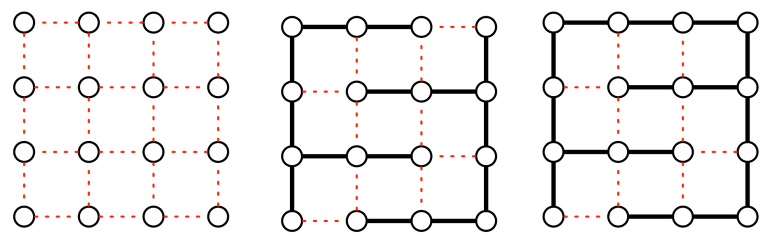
Problem: *ignores* 36 of the 128 components of the parameters and sufficient statistics in the example

Structured mean field harnesses an acyclic subgraph, but also takes into account all components

Question: how to choose the acyclic subgraph?

- Adding an edge in the subgraph can only increase quality
- But what is the impact on computational complexity?

Preview:



Complexity of computing gradient: $\mathcal{O}(n)$ $\mathcal{O}(n)$ $\mathcal{O}(n^3)$

First result: dichotomy in terms of a graph property, *v*-acyclic and *b*-acyclic subgraphs

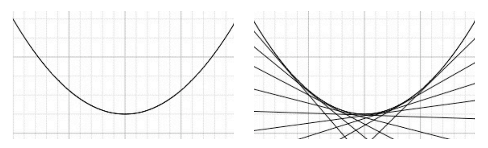
Second result: improved algorithm in the *b*-acyclic subgraph case

Background

Tool: Legendre-Fenchel transformation

$$f^*(x) = \sup\{\langle x, y \rangle - f(y) : y \in \mathrm{dom}(f)\}$$

Theorem: If f is convex and lower semi-continuous, $f = f^{**}$



Step 1: Express inference as a constrained optimization problem using convex duality:

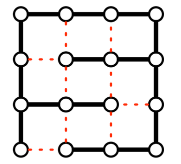
$$A(\boldsymbol{\theta}) = \sup\{\langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle - A^*(\boldsymbol{\mu}) : \boldsymbol{\mu} \in \mathscr{M}\}$$

$$\mathscr{M} = \{\boldsymbol{\mu} \in \mathbb{R}^D : \exists \boldsymbol{\theta} \in \Theta \text{ s.t. } \mathbb{E}[\boldsymbol{\phi}(\boldsymbol{X}_{\boldsymbol{\theta}})] = \boldsymbol{\mu}\}$$

Step 2: Relax the optimization problem using a subset of the initial exponential family (defined by a subgraph)

$\boldsymbol{Y}_{\boldsymbol{\omega}}$
Tractable r.v.

$\boldsymbol{\omega} \in \Xi \subseteq \mathbb{R}^d$
Tractable parameters


Subgraph

$$\hat{A}(\boldsymbol{\theta}) = \sup\{\langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle - A^*(\boldsymbol{\mu}) : \boldsymbol{\mu} \in \mathscr{M}_{\mathrm{MF}}\}$$

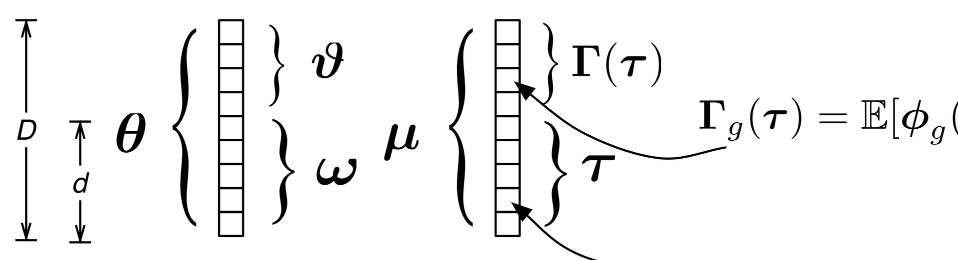
$$\mathscr{M}_{\mathrm{MF}} = \left\{\boldsymbol{\mu} \in \mathscr{M} : \exists \boldsymbol{\omega} \in \Xi \text{ s.t. } \mathbb{E}[\boldsymbol{\phi}(\boldsymbol{Y}_{\boldsymbol{\omega}})] = \boldsymbol{\mu}\right\}$$

Consequence: on $\boldsymbol{\mu} \in \mathscr{M}_{\mathrm{MF}}$, $A^*(\boldsymbol{\mu})$ is tractable

Step 3: Solve the simplified optimization problem

$$\hat{A}(\boldsymbol{\theta}) = \sup\{\langle \boldsymbol{\omega}, \boldsymbol{\tau} \rangle + \langle \boldsymbol{\vartheta}, \boldsymbol{\Gamma}(\boldsymbol{\tau}) \rangle - A_0^*(\boldsymbol{\tau}) : \boldsymbol{\tau} \in \mathscr{N}\}$$

realizable moments in the subgraph



$$\boldsymbol{\Gamma}_g(\boldsymbol{\tau}) = \mathbb{E}[\boldsymbol{\phi}_g(\boldsymbol{Y}_{\boldsymbol{\tau}})]$$

Necessary optimality condition:

$$0 = \boldsymbol{\omega} + J(\boldsymbol{\tau})\boldsymbol{\vartheta} - \nabla A_0^*(\boldsymbol{\tau})$$

$$\boldsymbol{\tau} = \underbrace{\nabla A_0}_{\text{Easy}}\big(\boldsymbol{\omega} + \underbrace{J(\boldsymbol{\tau})}_{?}\boldsymbol{\vartheta}\big)$$

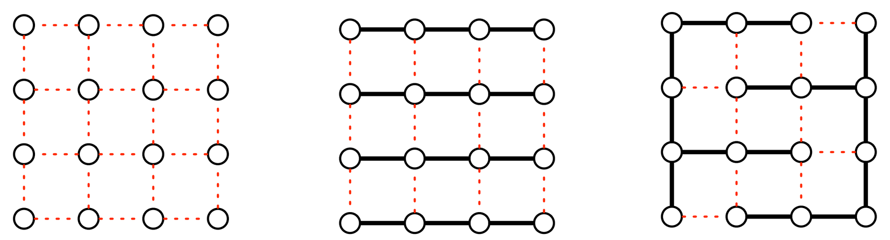
$$J = \left(\frac{\partial \Gamma_g}{\partial \tau_f}\right)_{f,g}$$

Dichotomy of tractable mean field subgraphs

Definition: an acyclic subgraph with edges $E' \subseteq E$ is ...

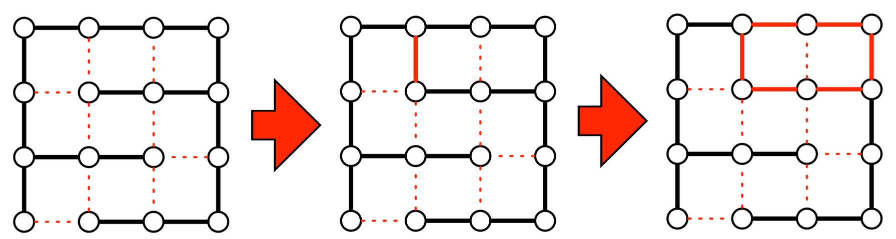
- *v*-acyclic, if for all $e \in E$, $E' \cup \{e\}$ is still acyclic
- *b*-acyclic, otherwise

Examples of *v*-acyclic graphs



Naive MF Factorial Maximum

Example of a *b*-acyclic graph



Bag of tricks

Properties of $A(\theta)$

$$\nabla A(\boldsymbol{\theta}) = \mathbb{E}[\boldsymbol{\phi}(\boldsymbol{X}_{\boldsymbol{\theta}})]$$
$$H(A(\boldsymbol{\theta})) = \mathbf{V}ar[\boldsymbol{\phi}(\boldsymbol{X}_{\boldsymbol{\theta}})]$$

$$\mathbf{V}ar \succeq 0 \Longrightarrow A(\theta) \text{ is convex}$$

$$\nabla A^* = \nabla A^{-1}$$

when the family is regular and minimal

Hammersley-Clifford Theorem

Consequence: if a, b belong to different cliques,

$$\boldsymbol{Y}_a \perp\!\!\!\perp \boldsymbol{Y}_b$$

Chain rule for Jacobian matrices

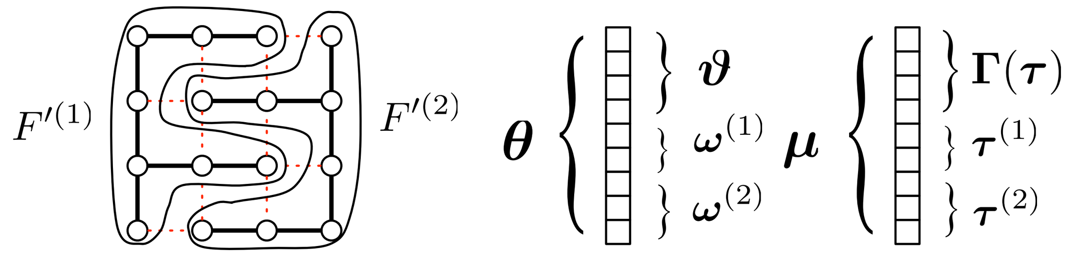
$$I = \left(\frac{\partial Z^{[g]}}{\partial \theta_h^{[g]}}\right)_{g,h}; \quad K = \left(\frac{\partial \theta_h^{[g]}}{\partial \tau_f}\right)_{h,f}$$

$$\Longrightarrow J = K^T I^T$$

v-acyclic subgraphs

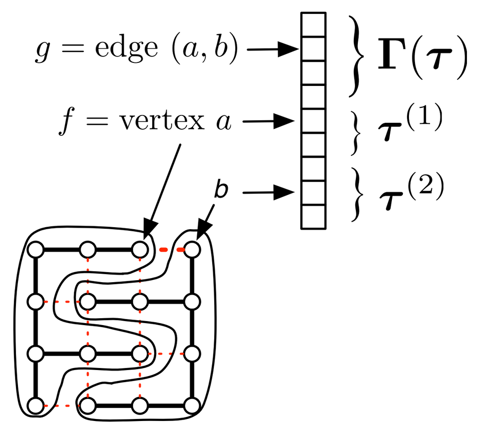
Connected component decomposition:

$$\boldsymbol{\tau} = \begin{pmatrix} \boldsymbol{\tau}^{(1)} \\ \boldsymbol{\tau}^{(2)} \end{pmatrix} = \begin{pmatrix} \nabla A_0^{(1)}\big(\boldsymbol{\omega}^{(1)} + J^{(1)}(\boldsymbol{\tau})\boldsymbol{\vartheta}\big) \\ \nabla A_0^{(2)}\big(\boldsymbol{\omega}^{(2)} + J^{(2)}(\boldsymbol{\tau})\boldsymbol{\vartheta}\big) \end{pmatrix}$$



Form of J :

$$\begin{aligned} J_{f,g}^{(1)}(\boldsymbol{\tau}) &= \frac{\partial}{\partial \tau_f}\mathbb{E}[\boldsymbol{\phi}_g(\boldsymbol{Y}_{\boldsymbol{\tau}})] \\ &= \frac{\partial}{\partial \tau_f}\mathbb{P}(Y_a = s, Y_b = t) \\ &= \frac{\partial}{\partial \tau_f}\mathbb{P}(Y_a = s)\mathbb{P}(Y_b = t) \\ &= \frac{\partial}{\partial \tau_f}\tau_{a,s}\tau_{b,t} = \boldsymbol{\tau}_{b,t} \end{aligned}$$



Relation to block Gibbs sampling

$$\boldsymbol{X}_t^{(2)} \Big| \boldsymbol{X}_{t-1} \sim \mathrm{MRF}\left(\boldsymbol{\omega}^{(2)} + B^{(2)}(\boldsymbol{X}_{t-1})\boldsymbol{\vartheta}\right)$$

$J(\boldsymbol{\tau})$

0.1	0.3			
0.9		0.8		
			0.4	0.9
	0.2		0.6	
	0.5	0.2		0.1

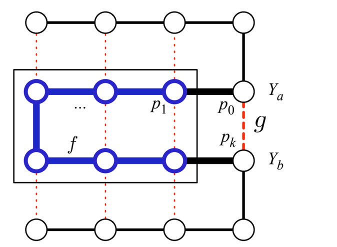
$B(\boldsymbol{X}_t)$

0	1			
1		1		
			0	1
	0		1	
	0	0		0

b-acyclic subgraphs

Form of J :

$$\begin{aligned} J_{f,g}(\boldsymbol{\tau}) &= \frac{\partial}{\partial \tau_f}\mathbb{P}(Y_a = s, Y_b = t) \\ &= \frac{\partial}{\partial \tau_f}\sum_{y_1 \in \mathcal{X}} \cdots \sum_{y_{k-1} \in \mathcal{X}} \mathbb{P}(Y_{p_i} = y_i, \forall i \in \{1, \dots k\}) \\ &= \frac{\partial}{\partial \tau_f}\mathbb{P}(Y_a = s)\sum_{y_1 \in \mathcal{X}} \mathbb{P}(Y_{p_1} = y_1 | Y_{p_0} = y_0) \sum_{y_2 \in \mathcal{X}} \cdots \\ &= \frac{\partial}{\partial \tau_f}\tau_{a,s}\sum_{y_1 \in \mathcal{X}} \frac{\tau_{(p_0,p_1),(y_0,y_1)}}{\tau_{p_0,y_0}} \sum_{y_2 \in \mathcal{X}} \cdots \end{aligned}$$



Technique: auxiliary exponential families

For fixed g, construct an exponential families such that its partition function satisfies:

$$\begin{aligned} Z^{[g]}(\boldsymbol{\theta}^{[g]}) &= \sum_{\boldsymbol{x} \in \mathcal{X}^{k-1}} \exp\{\langle \boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\theta}^{[g]} \rangle\} \\ &= \Big(\sum_{s'} \tau_{(a,p_1),(s,s')}\Big) \sum_{y_1 \in \mathcal{X}} \frac{\tau_{(p_0,p_1),(y_0,y_1)}}{\big(\sum_{s'} \tau_{(p_0,p_1),(y_0,s')}\big)} \\ &\quad \times \sum_{y_2 \in \mathcal{X}} \cdots \sum_{y_{k-1} \in \mathcal{X}} \frac{\tau_{(p_{k-2},p_{k-1}),(y_{k-2},y_{k-1})}}{\big(\sum_{s'} \tau_{(p_{k-2},p_{k-1}),(y_{k-2},s')}\big)} \\ &\quad \times \frac{\tau_{(p_{k-1},p_k),(y_{k-1},y_k)}}{\big(\sum_{s'} \tau_{(p_{k-1},p_k),(y_{k-1},s')}\big)} \end{aligned}$$

Why? We can get all the derivatives of the log-partition function in one shot using sum-product

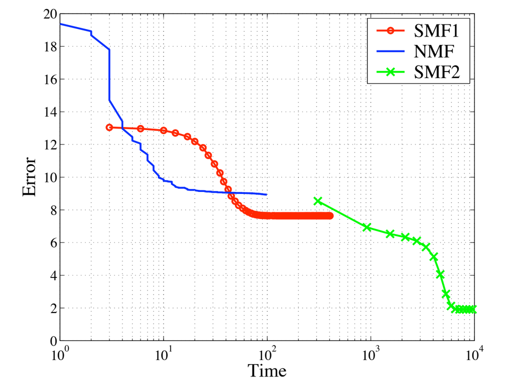
$$\frac{\partial Z^{[g]}}{\partial \theta_h^{[g]}} = Z^{[g]} \times \frac{\partial A^{[g]}}{\partial \theta_h^{[g]}} = Z^{[g]} \times \mu_h^{[g]}$$

How can we get the partial derivative with respect to $\boldsymbol{\tau}$?

Experiments

Adding edges improves the quality of the approximation

Using a *b*-acyclic subgraph is significantly more expensive



References

D. Barber and W. Wiegerinck. Tractable variational structures for approximating graphical models. In *Advances in Neural Information Processing Systems*, pages 183–189, Cambridge, MA, 1999. MIT Press.

D. M. Blei and M. I. Jordan. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1:121–144, 2005.

N. De Freitas, P. Hojen-Sorensen, M. I. Jordan, and S. Russell. Variational MCMC. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, San Mateo, CA, 2001. Morgan Kaufmann.

D. Geiger, C. Meek, and Y. Wexler. A variational inference procedure allowing internal structure for overlapping clusters and deterministic constraints. *Journal of Artificial Intelligence Research*, 27:1–23, 2006.

A. Globerson and T. Jaakkola. Approximate inference using planar graph decomposition. In *Advances in Neural Information Processing Systems*, Cambridge, MA, 2006. MIT Press.

G. Hua and Y. Wu. Sequential mean field variational analysis of structured deformable shapes. *Computer Vision and Image Understanding*, 101:87–99, 2006.

J. B. Lasserre. Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization*, 11:796–817, 2000.

T. Minka and Y. Qi. Tree-structured approximations by expectation propagation. In *Advances in Neural Information Processing Systems*, Cambridge, MA, 2003. MIT Press.

C. Peterson and J. R. Anderson. A mean field theory learning algorithm for neural networks. *Complex Systems*, 1: 995–1019, 1987.

L. K. Saul and M. I. Jordan. Exploiting tractable substructures in intractable networks. In *Advances in Neural Information Processing Systems 8*, pages 486–492, Cambridge, MA, 1996. MIT Press.

M. J. Wainwright and M. I. Jordan. Variational inference in graphical models: The view from the marginal polytope. In *Forty-first Annual Allerton Conference on Communication, Control, and Computing*, 2003.

M. J. Wainwright and M. I. Jordan. Log-determinant relaxation for approximate inference in discrete Markov random fields. *IEEE Transactions on Signal Processing*, 54:2099–2109, 2006.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1:1–305, 2008.

W. Wiegerinck. Variational approximations between mean field theory and the junction tree algorithm. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, pages 626–633, San Mateo, CA, 2000. Morgan Kaufmann.

J. S. Yedidia, W. T. Freeman, and Y. Weiss. Generalized belief propagation. In *Advances in Neural Information Processing Systems*, pages 689–695, Cambridge, MA, 2001. MIT Press.