Discrete Markov Random Fields
the Inference story

Pradeep Ravikumar

Graphical Models, The History

How to model stochastic processes of the world?

I want to model the world, and I like graphs...

History

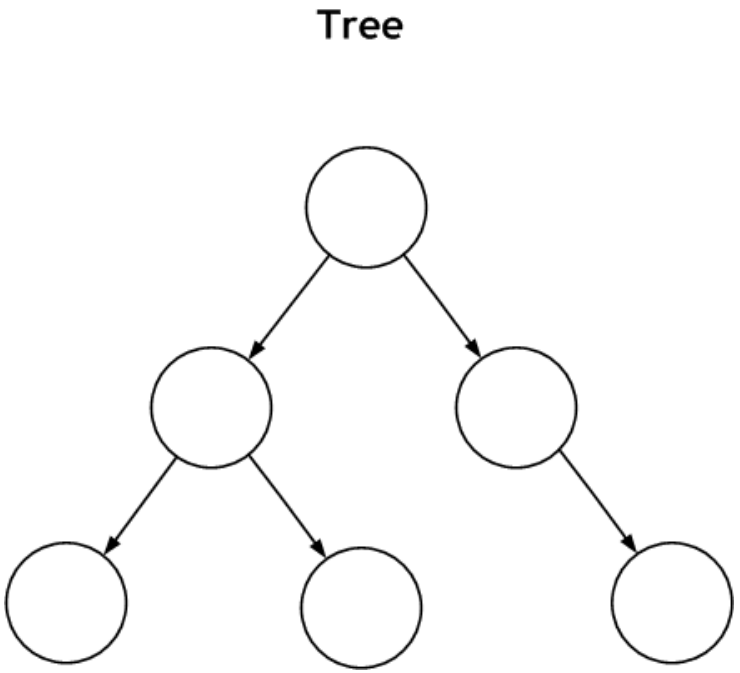
Mid to Late Twentieth Century

Pioneering work of Conspiracy Theorists

The System, it is all **connected**...

History

Late Twentieth Century: people realize that existing scientific literature offers a marriage between probability theory and graph theory – which can be used to model the world.



History

Common Misconception: Called graphical models after Grafici Modeles, a sculptor protege of Da Vinci.

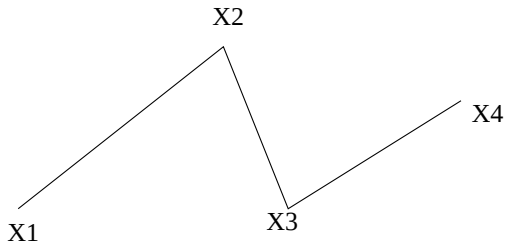
Called Graphical Models because it models stochastic systems using graphs.

History

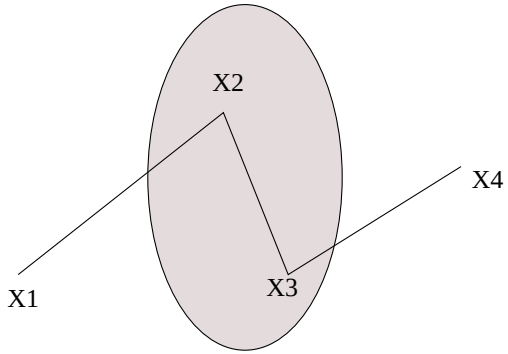
Common Misconception: Called graphical models after Grafici Modeles, a sculptor protege of Da Vinci.

Called Graphical Models because it models stochastic systems using graphs.

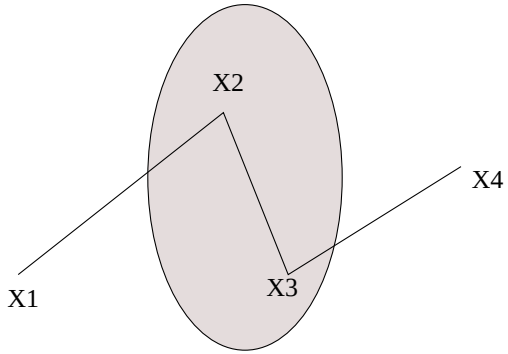
Graphical Models



Graphical Models

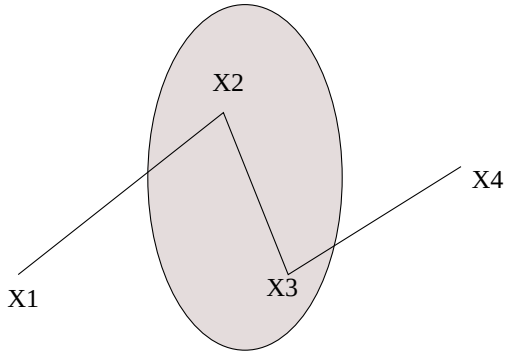


Graphical Models



Separating Set $\sim (X_2, X_3)$ disconnects X_1 and X_4

Graphical Models



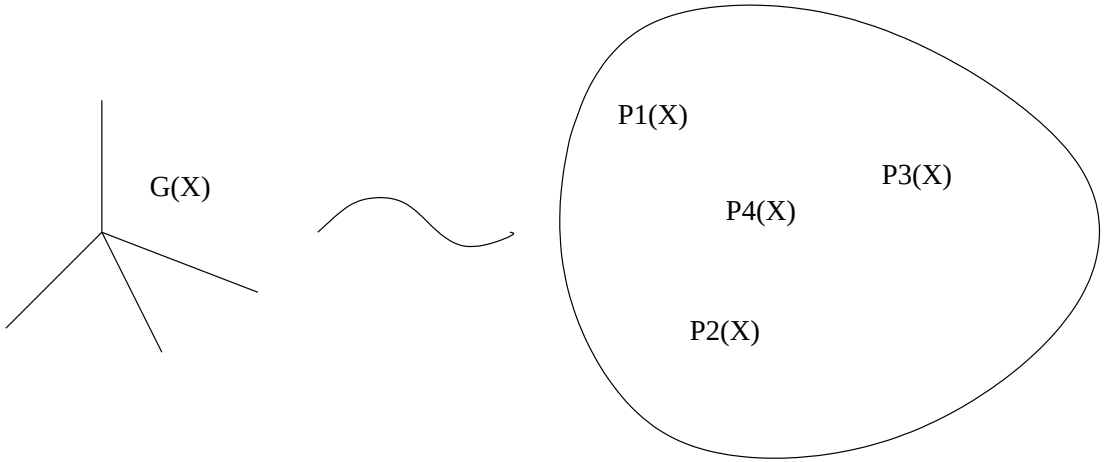
Separating Set $\sim (X_2, X_3)$ disconnects X_1 and X_4

Global Markov Property $\sim X_1 \perp X_4 \,|\, (X_2,\, X_3)$

Graphical Models

$\mathcal{MP}(G) \sim$ Set of all Markov properties, by ranging over separating sets of G.

$\mathcal{P}$ represented by $G \sim \mathcal{P}$ satisfies $\mathcal{MP}(G)$

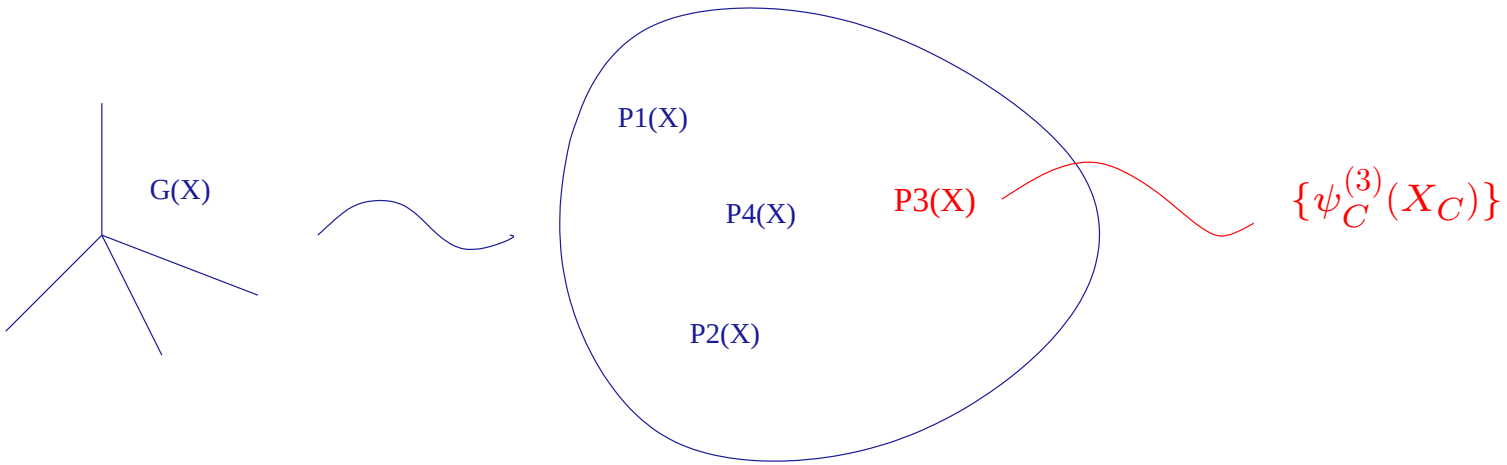


Hammersley and Clifford Theorem

Positive $\mathcal{P}$ over X satisfies $\mathcal{MP}(G)$ iff $\mathcal{P}$ factorizes according to cliques $\mathcal{C}$ in G,

$$\mathcal{P}(X) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(X_C)$$

Specific member of family specified by weights over cliques.



Exponential Family

$$
\begin{aligned}
p(X) &= \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(X_C) \\
&= \exp\Big(\sum_{C \in \mathcal{C}} \log \psi_C(X_C) - \log Z\Big)
\end{aligned}
$$

Exponential family: $p(X;\theta) = \exp\left(\sum_{\alpha \in \mathcal{C}} \theta_\alpha \phi_\alpha(X) - \Psi(\theta)\right)$

- $\{\phi_\alpha\} \sim$ features
- $\{\theta_\alpha\} \sim$ parameters
- $\Psi(\theta) \sim$ log partition function

Inference

Answering queries about the graphical model probability distribution.

Inference

For undirected model $p(x;\theta) = \exp\left(\sum_{\alpha\in I}\theta_\alpha\phi_\alpha(x) - \Psi(\theta)\right)$ key inference problems are:

- compute log partition function (normalization constant) $\Psi(\theta)$
- marginals $p(x_A) = \sum_{x_v, v\notin A} p(x)$
- most probable configurations $x^* = \arg\max_x p(x \mid x_L)$

These problems are intractable in full generality.

Log Partition Function

$$Z = \log \sum_{x} \prod_{\alpha \in I} \psi_\alpha(x_\alpha)$$

Variable Elimination

$$Z = \log \sum_{x} \prod_{\alpha \in I} \psi_\alpha(x_\alpha)$$

$$\begin{aligned}
&\sum_{x} \quad \prod_{\alpha} \psi_\alpha(x_\alpha) \\
&= \sum_{\{x_{j\neq i}\}} \sum_{x_i} \prod_{\alpha} \psi_\alpha(x_\alpha) \\
&= \sum_{\{x_{j\neq i}\}} \prod_{\alpha \in C_{\setminus i}} \psi_\alpha(x_\alpha) \sum_{x_i} \prod_{\alpha \in C_i} \psi_\alpha(x_\alpha) \\
&= \sum_{\{x_{j\neq i}\}} \prod_{\alpha \in C_{\setminus i}} \psi_\alpha(x_\alpha) g(x_{j\neq i})
\end{aligned}$$

Variable Elimination

$$Z = \sum_{\{x_{j\neq i}\}} \prod_{\alpha \in C_{\setminus i}} \psi_\alpha(x_\alpha) g(x_{j\neq i})$$

Continue to "eliminate" other variables x_j.

Is this a linear time method then?

Variable Elimination

$$Z = \sum_{\{x_{j\neq i}\}} \prod_{\alpha \in C_{\setminus i}} \psi_\alpha(x_\alpha) g(x_{j\neq i})$$

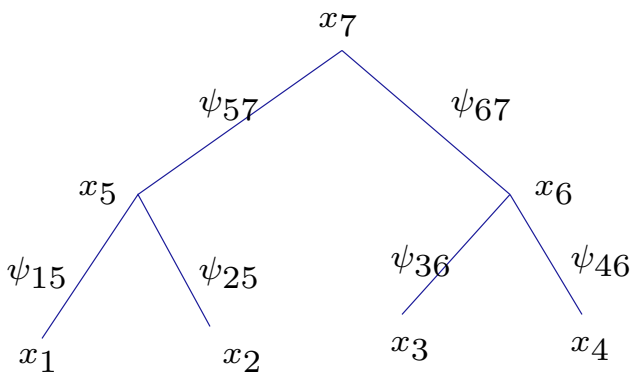
Continue to "eliminate" other variables x_j.

Is this a linear time method then?

$g(x_{j\neq i})$ depends on variables j which share a factor with i.

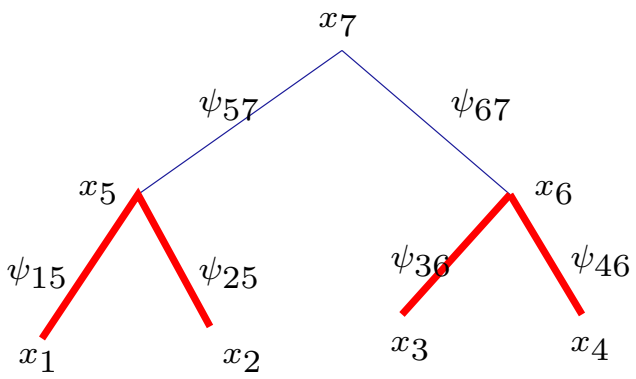
Variable Elimination

$$Z = \log \sum_{x} \prod_{\alpha \in I} \psi_\alpha(x_\alpha)$$



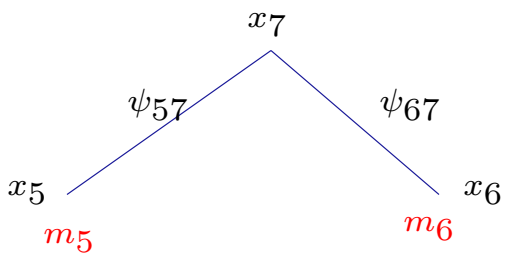
Variable Elimination

$$Z = \log \sum_{x} \prod_{\alpha \in I} \psi_\alpha(x_\alpha)$$



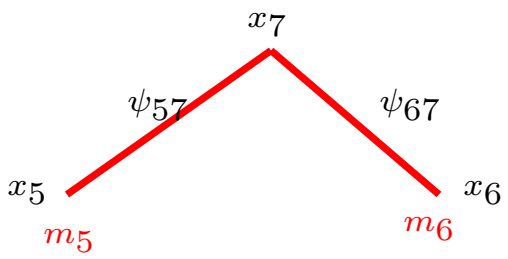
Variable Elimination

$$Z = \log \sum_{x} \prod_{\alpha \in I} \psi_\alpha(x_\alpha)$$



Variable Elimination

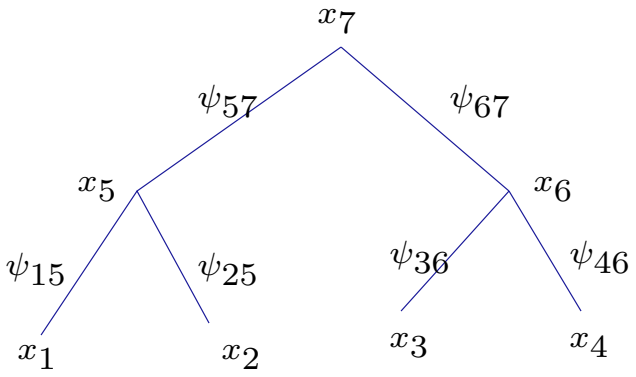
$$Z = \log \sum_{x} \prod_{\alpha \in I} \psi_\alpha(x_\alpha)$$



Variable Elimination

$$Z = \log \sum_{x} \prod_{\alpha \in I} \psi_\alpha(x_\alpha)$$

Exponential in tree-width.



$$\sum_{x_7} \quad \sum_{x_5} \psi_{57} \left(\sum_{x_1} \psi_{15} \sum_{x_2} \psi_{25}\right)$$
$$\sum_{x_6} \psi_{67} \left(\sum_{3} \psi_{36} \sum_{x_4} \psi_{46}\right)$$

Inference

$$p(x;\theta) = \exp(\theta^\top \phi(x) - A(\theta))$$

$$A(\theta) \sim \text{log partition function}$$

Inference

$$
\begin{aligned}
A(\theta) &= \log \sum_x \exp(\theta^\top \phi(x)) \\
&\leq B(\theta, \lambda) \\
&\geq C(\theta, \lambda)
\end{aligned}
$$

$$
\lambda \sim \text{“variational” parameter}
$$

$$
\begin{aligned}
A(\theta) &\leq \inf_\lambda B(\theta, \lambda) \\
&\geq \sup_\lambda C(\theta, \lambda)
\end{aligned}
$$

Summing over configurations $\longrightarrow$ Optimization!

Inference

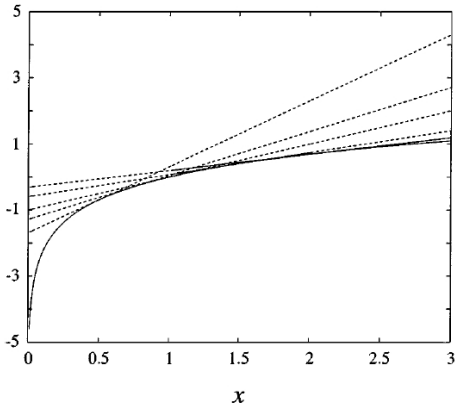
But... (there's always a but!)
is there a principled way to obtain "parametrized" bounds $B(\theta, \lambda)$ and $C(\theta, \lambda)$?

Fenchel Duality

$f(x) \sim$ concave function

Define $f^*(\lambda) = \min_x\{\lambda^\top x - f(x)\}$.

$\Longrightarrow f^{'}(x_\lambda) = \lambda$, Slope of tangent at x_λ is λ

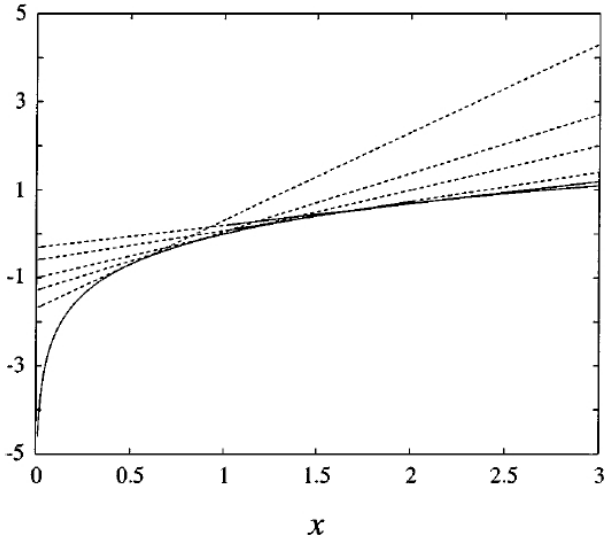


Tangent $\sim \lambda^\top x - (Intercept)$

$\Longrightarrow f(x_\lambda) = \lambda^\top x_\lambda - (Intercept)$

$\Longrightarrow f^*(\lambda) \sim$ Intercept of line with slope λ tangent to $f(x)$.

Fenchel Duality



Tangent $\sim \lambda^\top x - f^*(\lambda)$

$f(x) = \min_\lambda \{\lambda^\top x - f^*(\lambda)\}$

Thus, $g(x, \lambda) = \lambda^\top x - f^*(\lambda)$ is an upper bound of $f(x)$!

Fenchel Duality

Let us apply fenchel duality to the log partition function!

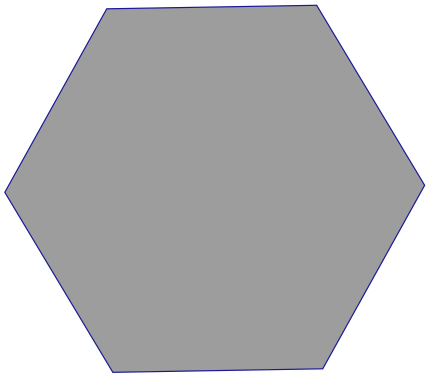
$$
\begin{aligned}
A(\theta) &\sim \text{convex} \\
A^*(\mu) &= \sup_{\theta}(\theta^\top \mu - A(\theta)) \\
A(\theta) &= \sup_{\mu}(\theta^\top \mu - A^*(\mu))
\end{aligned}
$$

Log Partition Function

Define the "marginal polytope"

$$\mathcal{M} = \{\mu \in \mathbb{R}^d | \exists p(.) \text{ s.t. } \sum_x \phi(x)p(x) = \mu\}$$

$$\mathcal{M} \sim \text{Convex hull of } \{\phi(x)\}$$



Mean parameter mapping

Consider the mapping $\wedge : \Theta \to \mathcal{M}$,

$$\wedge(\theta) := E_\theta[\phi(x)] = \sum_x \phi(x) p(x;\theta)$$

The mapping associates θ to "mean parameters" $\mu := \wedge(\theta) \in \mathcal{M}$.

Conversely, for μ in $\mathrm{Int}(\mathcal{M})$, $\exists \theta = \wedge^{-1}(\mu)$ (unique if exponential family is minimal)

Partition function conjugates

$$
\begin{aligned}
A^*(\mu) &= \sup_{\theta}(\theta^\top \mu - A(\theta)) \\
A(\theta) &= \sup_{\mu}(\theta^\top \mu - A^*(\mu))
\end{aligned}
$$

Optimal parameters given by,

$$
\begin{aligned}
\theta_\mu &= \wedge^{-1}(\mu) \\
\mu_\theta &= \wedge(\theta)
\end{aligned}
$$

Partition function conjugate

Properties of the fenchel conjugate, $A^*(\mu) = \sup_\theta(\theta^\top \mu - A(\theta))$

- $A^*(\mu)$ is finite only for $\mu \in \mathcal{M}$.
- $A^*(\mu)$ is the entropy of graphical model distribution with "mean parameters" μ, or equivalently with parameters $\wedge^{-1}(\mu)$!

Partition Function

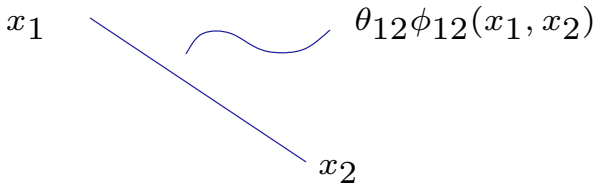
$$A(\theta) = \sup_{\mu \in \mathcal{M}} \theta^\top \mu - A^*(\mu)$$

"Hardness" is due to two bottlenecks

- $\mathcal{M}$: a polytope with exponentially many vertices and no compact representation
- $A^*(\mu)$: entropy computation

Approximate either or both!

Pairwise Graphical Models



Overcomplete potentials:

$$\mathcal{I}_j(x_s) = \begin{cases} 1 & x_s = j \\ 0 & \text{otherwise} \end{cases}$$

$$\mathcal{I}_{j,k}(x_s, x_t) = \begin{cases} 1 & x_s = j \text{ and } x_t = k \\ 0 & \text{otherwise}. \end{cases}$$

$$p(x|\theta) = \exp\left(\sum_{s,j} \theta_{s;j}\mathcal{I}_j(x_s) + \sum_{s,t;j,k} \theta_{s,t;j,k}\mathcal{I}_{j,k}(x_s, x_t) - \Psi(\theta)\right)$$

Overcomplete Representation; Mean Parameters

$$\begin{aligned} \mu_{s;j} &:= E_\theta[\mathcal{I}_j(x_s)] = p(x_s = j; \theta) \\ \mu_{s,t;j,k} &:= E_\theta[\mathcal{I}_{j,k}(x_s, x_t)] = p(x_s = j, x_t = k; \theta) \end{aligned}$$

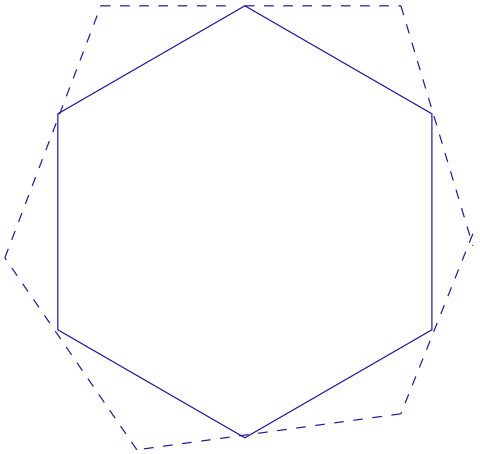
Mean parameters are marginals!

Define the following functional forms,

$$\begin{aligned} \mu_s(x_s) &= \sum_j \mu_{s;j} \mathcal{I}_j(x_s) \\ \mu_{st}(x_s, x_t) &= \sum_{j,k} \mu_{s,t;j,k} \mathcal{I}_{jk}(x_s, x_t) \end{aligned}$$

Outer Polytope Approximations

$$\mathrm{LOCAL}(G) := \{\mu \geq 0 | \sum_{x_s} \mu_s(x_s) = 1, \sum_{x_t} \mu_{st}(x_s, x_t) = \mu_s(x_s)\}$$

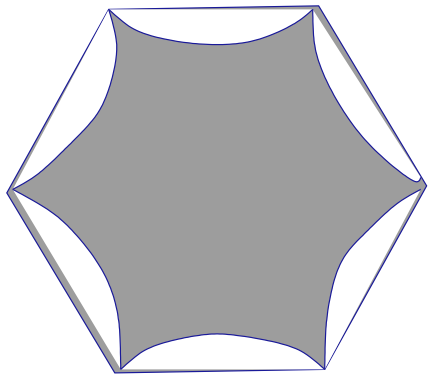


Inner Polytope Approximations

For the given graph G and a subgraph H, let

$$\mathcal{E}(H) = \{\theta' \,|\, \theta'_{st} = \theta_{st}\, 1_{(s,t)\in H}\}$$

$$\begin{aligned} \mathcal{M}(G;H) &= \{\mu \,|\, \mu = E_\theta[\phi(x)] \text{ for some } \theta \in \mathcal{E}(H)\}\,. \\ \mathcal{M}(G;H) &\subseteq \mathcal{M}(G) \end{aligned}$$



Entropy Approximations

Tree-structured distributions,

$$p(x;\mu) = \prod_s \mu_s(x_s) \prod_{(s,t)\in E} \frac{\mu_{st}(x_s,x_t)}{\mu_s \mu_t}$$

Define,

$$\begin{aligned} H_s(\mu_s) &:= \sum_{x_s} \mu_s(x_s) \log \mu_s(x_s) \\ I_{st}(\mu_{st}) &:= \sum_{x_s,x_t} \mu_{st}(x_s,x_t) \log \mu_{st}(x_s,x_t) \end{aligned}$$

Tree-structured Entropy,

$$A^*_{tree}(\mu) = -\sum_{s \in V} H_s(\mu_s) + \sum_{(s,t) \in E} I_{st}(\mu_{st})$$

Compact representation; can be used as an approximation.

Approximate Inference Techniques

Belief Propagation – Polytope $\sim LOCAL(G)$, Entropy $\sim$ Tree-structured entropy!

Structured Mean Field – Polytope $\sim \mathcal{M}(G; H)$, Entropy $\sim$ H-structured entropy
Mean Field – $H = H_0$, completely independent graph

Divergence Measure View

Given: $p(x;\theta) \propto \exp(\theta^\top \phi(x))$

Would like a more "manageable" surrogate distribution, $q \in \mathcal{Q}$,

$$\min_{q \in \mathcal{Q}} D(q(x)||p(x;\theta))$$

Divergence Measure View

$$\min_{q \in \mathcal{Q}} D(q(x)||p(x;\theta))$$

$D(q||p) = KL(q||p) \sim$ Structured Mean Field, Belief Propagation

$D(p||q) = KL(p||q) \sim$ Expectation Propagation

(look out for talk on Continuous Markov Random Fields!)

Typically approximate KL measure with "energy approximations"

(Bethe free energy, Kikuchi free energy)

(Ravikumar,Lafferty 05; Preconditioner Approximations)

Optimizing for a minimax criterion reduces task to a generalized linear systems

problem!

Bounds on event probabilities

Doctor: So what is the lower bound on the diagnosis probability?

Graphical Model: I don't know, but here is an "approximate" value.

Doctor: $=($

Can we get upper and lower bounds on $p(X \in C; \theta)$ (instead of just "approximate" values)?

Bounds on event probabilities

Classical Chernoff Bounds give useful estimates for **i.i.d.** random variables.

Can they be extended to graphical models?

[Ravikumar, Lafferty 04; Variational Chernoff Bounds]

Classical Chernoff Bounds

$p_\theta(X \geq u) \leq \frac{E_\theta(X)}{u}$ *Markov Inequality*

$p_\theta(X \geq u) = p_\theta(e^{\lambda X} \geq e^{\lambda u}) \leq E_\theta[e^{\lambda(X-u)}]$

From this it follows that:

$\log\, p_\theta(X \geq u) \leq \inf_{\lambda \geq 0}(-\lambda u + \log\, E_\theta[e^{\lambda X}])$

Bounds on cumulant function $E_\theta[e^{\lambda X}]$ yield the standard Chernoff Bounds.

Classical Chernoff Bounds

$p_\theta(X \geq u) \leq \frac{E_\theta(X)}{u}$ *Markov Inequality*

$p_\theta(X \geq u) = p_\theta(e^{\lambda X} \geq e^{\lambda u}) \leq E_\theta[e^{\lambda(X-u)}]$

From this it follows that:

$\log\, p_\theta(X \geq u) \leq \inf_{\lambda \geq 0}(-\lambda u + \log\, E_\theta[e^{\lambda X}])$

Bounds on cumulant function $E_\theta[e^{\lambda X}]$ yield the standard Chernoff Bounds.

Classical Chernoff Bounds

$p_\theta(X \geq u) \leq \frac{E_\theta(X)}{u}$ *Markov Inequality*

$p_\theta(X \geq u) = p_\theta(e^{\lambda X} \geq e^{\lambda u}) \leq E_\theta[e^{\lambda(X-u)}]$

From this it follows that:

$\log\, p_\theta(X \geq u) \leq \inf_{\lambda \geq 0}(-\lambda u + \log\, E_\theta[e^{\lambda X}])$

Bounds on cumulant function $E_\theta[e^{\lambda X}]$ yield the standard Chernoff Bounds.

Generalized Chernoff Bounds

Event: $X \in C$

$$\mathcal{I}_C(X) = \begin{cases} 1 & X \in C \\ 0 & \text{otherwise} \end{cases}$$

$\mathcal{I}_C(X) \leq f_\lambda$

$p_\theta(X \in C) \leq E_\theta[f_\lambda]$

Generalized Chernoff Bounds

Event: $X \in C$

$$\mathcal{I}_C(X) = \begin{cases} 1 & X \in C \\ 0 & \text{otherwise} \end{cases}$$

$$\mathcal{I}_C(X) \leq f_\lambda$$

$$p_\theta(X \in C) \leq E_\theta[f_\lambda]$$

Graphical Model Chernoff Bounds

With $f_\lambda = \exp(\langle \lambda, x \rangle + u)$, we get:

$$\log\, p_\theta(X \in C) \leq \inf_\lambda (S_C(-\lambda) + \log\, E_\theta[e^{<\lambda, x>}])$$

where $S_C(\lambda) = \sup_{x \in C} \langle x, \lambda \rangle$ is the support function of the set C.

For an exponential model with sufficient statistic $\phi(x)$, the above becomes:

$$\log\, p_\theta(X \in C) \leq \inf_\lambda (S_{C,\phi}(-\lambda) + \Phi(\theta + \lambda) - \Phi(\theta))$$

where Φ is the log-partition function and $S_{C,\phi}(\lambda) = \sup_{x \in C} \langle \phi(x), \lambda \rangle$

Graphical Model Chernoff Bounds

With $f_\lambda = \exp(\langle \lambda, x \rangle + u)$, we get:

$$\log\, p_\theta(X \in C) \leq \inf_\lambda (S_C(-\lambda) + \log\, E_\theta[e^{<\lambda, x>}])$$

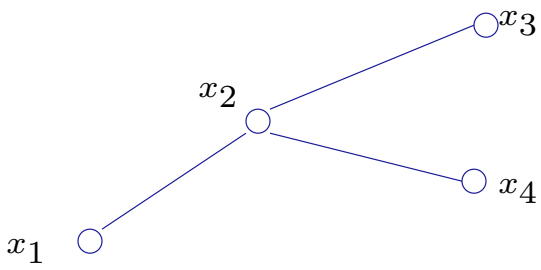
where $S_C(\lambda) = \sup_{x \in C} \langle x, \lambda \rangle$ is the support function of the set C.

For an exponential model with sufficient statistic $\phi(x)$, the above becomes:

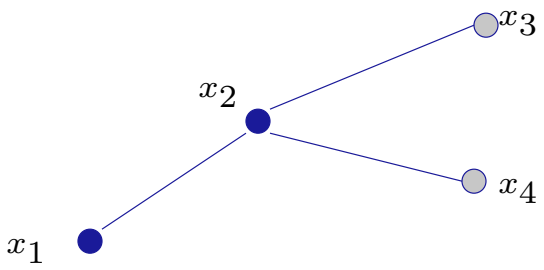
$$\log\, p_\theta(X \in C) \leq \inf_\lambda (S_{C,\phi}(-\lambda) + \Phi(\theta + \lambda) - \Phi(\theta))$$

where Φ is the log-partition function and $S_{C,\phi}(\lambda) = \sup_{x \in C} \langle \phi(x), \lambda \rangle$

MAP Estimation

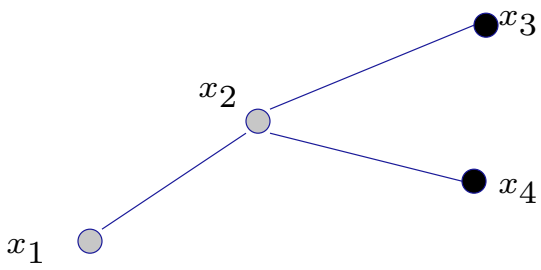


MAP Estimation



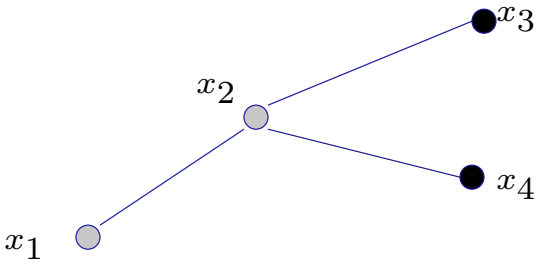
$$\text{PROB} = 0.01$$

MAP Estimation



$$\text{PROB} = 0.2$$

MAP Estimation

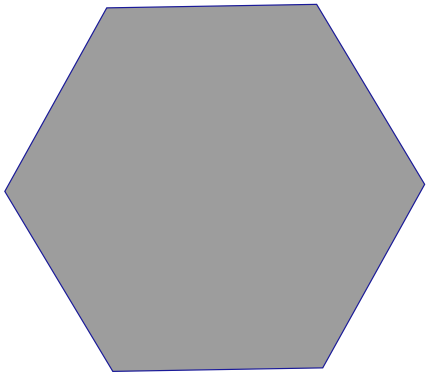


$$\text{PROB} = 0.2$$

Most Probable Configuration?

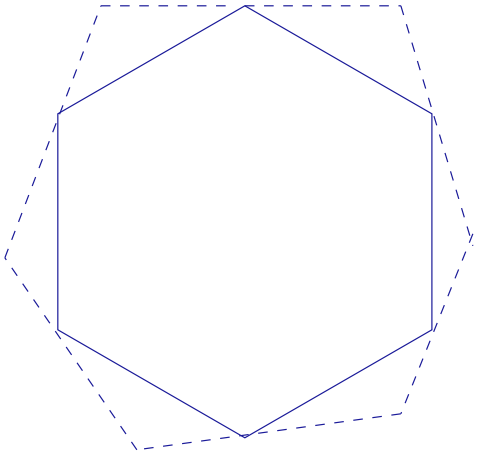
Polytope View

$$\mu^* = \max_x \theta^\top \phi(x) = \sup_{\mu \in \mathcal{M}} \theta^\top \mu$$



Outer Polytope Relaxations

$$\mathrm{LOCAL}(G) := \{\mu \geq 0 | \sum_{x_s} \mu_s(x_s) = 1, \sum_{x_t} \mu_{st}(x_s, x_t) = \mu_s(x_s)\}$$



Outer Polytope Relaxations

$$\begin{aligned} & \sup_{\mu \in \mathcal{M}(G)} \quad \theta^\top \mu \\ \leq & \sup_{\mu \in \mathrm{LOCAL}(G)} \quad \theta^\top \mu \end{aligned}$$

A Linear Program!

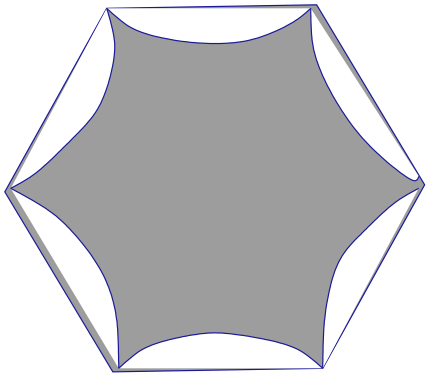
(Chekuri, Khanna, Naor, Zosin 05; LP Formulation for Metric Labeling)

(Wainwright, Jaakkola, Willsky 05; Tree-reweighted Max-Product, Dual of LP)

Inner Polytope Approximations

If $M_I \subset \mathcal{M}$ is any subset of the marginal polytope that includes all of the vertices,

$$\mu^* = \max_x \langle \theta, \phi \rangle = \sup_{\mu \in M_I} \langle \theta, \mu \rangle$$

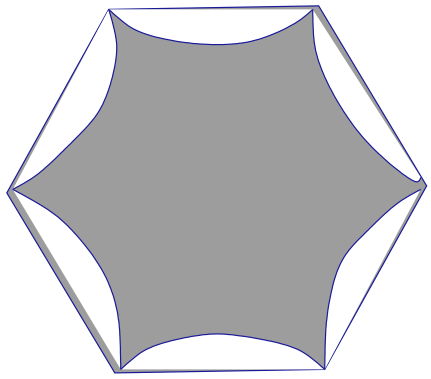


Inner Polytope Approximations

For the given graph G and a subgraph H, let

$$\mathcal{E}(H) = \{\theta' \,|\, \theta'_{st} = \theta_{st}\, 1_{(s,t)\in H}\}$$

$$\begin{aligned} \mathcal{M}(G;H) &= \{\mu \,|\, \mu = E_\theta[\phi(x)] \text{ for some } \theta \in \mathcal{E}(H)\}\,. \\ \mathcal{M}(G;H) &\subseteq \mathcal{M}(G) \end{aligned}$$



Inner Polytope Approximations

Mean Field parameters,

$$\mathcal{M}(G; H_0) \quad = \quad \{\mu(s;j), \mu(s,j;t,k) \,|\, 0 \leq \mu(s;j) \leq 1, \mu(s,j;t,k) = \mu(s;j)\mu(t;k)\}$$

Mean Field Relaxation,

$$\begin{aligned}
&\sup_{\mu \in \mathcal{M}(G;H_0)} \langle \theta, \mu \rangle \\
&= \sup_{\mu \in \mathcal{M}(G;H_0)} \sum_{s;j} \theta_{s;j}\mu(s;j) + \sum_{st;jk} \theta_{s,j;t,k}\mu(s,j;t,k) \\
&= \sup_{\mu \in \mathcal{M}(G;H_0)} \sum_{s;j} \theta_{s;j}\mu(s;j) + \sum_{st;jk} \theta_{s,j;t,k}\mu(s;j)\mu(t;k)
\end{aligned}$$

Quadratic Program!

(Ravikumar, Lafferty 06; Quadratic Relaxations for Metric Labeling and MAP in MRFs)

References

- Martin. J. Wainwright and Michael I. Jordan (2003). Graphical models, exponential families, and variational inference.
- M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul (1999). An introduction to variational methods for graphical models.
- Pradeep Ravikumar, John Lafferty (2005). Preconditioner Approximations for Probabilistic Graphical Models.
- Pradeep Ravikumar, John Lafferty (2004). Variational Chernoff Bounds for Graphical Models.
- Chekuri, C., Khanna, S., Naor, J., Zosin, L. (2005). A linear programming formulation and approximation algorithms for the metric labeling problem.
- Pradeep Ravikumar, John Lafferty (2006). Quadratic Programming Relaxations for Metric Labeling and Markov Random Field MAP Estimation.