UNIVERSITY OF MINNESOTA
Driven to Discover℠

Markov Random Fields: Inference and Estimation

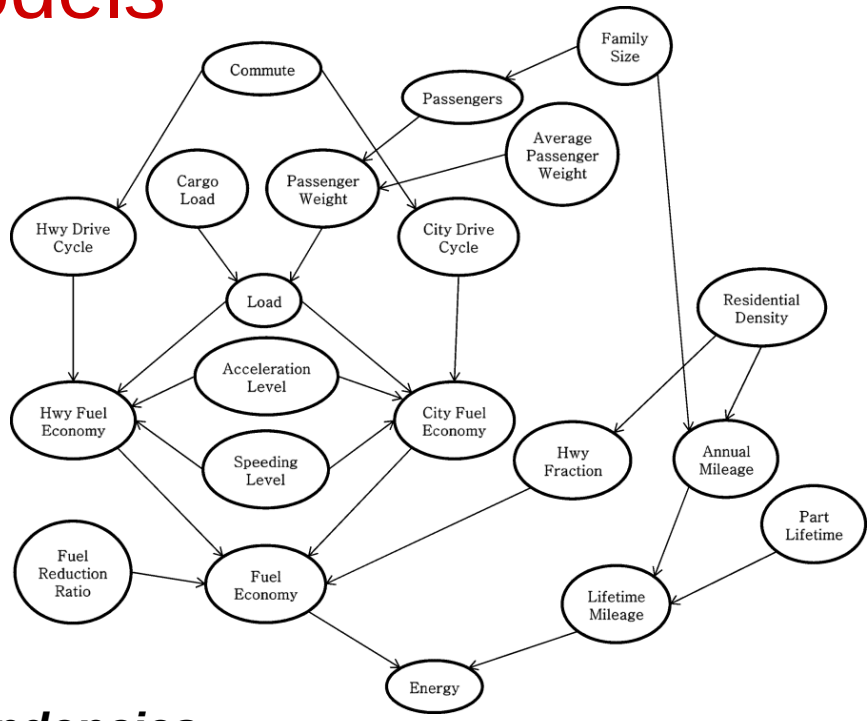
SPiNCOM reading group
April 24th , 2017

Dimitris Berberidis

Ack: Juan-Andres Bazerque

Probabilistic graphical models

- Set of random variables $\mathbf{x} = [x_1 x_2 \ldots x_N]$
  - Graph $\mathcal{G}$ represents joint $p(\mathbf{x})$
  - Nodes correspond to random variables
  - Edges imply relations between rv's
- Some applications
  - Speech recognition, computer vision
  - Decoding
  - Gene reg. networks, disease diagnosis



- Key idea: Graph models ***conditional independencies***
  - Two main tasks: Inference and Estimation

Inference: Given observed $\mathbf{x}_s$, obtain (marginal) conditionals $p(x_i|\mathbf{x}_s) \quad , \forall i \in \mathcal{S}^c$

Estimation: Given samples $\{\mathbf{x}^{(t)}\}_{t=1}^{T}$ estimate $\mathcal{G}$ (and thus $p(\mathbf{x})$)

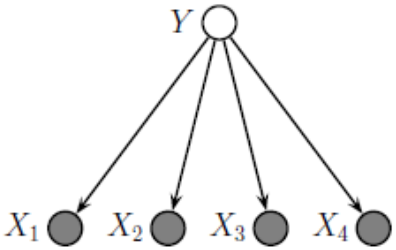
Roadmap

- Bayesian networks basics
- Markov Random Fields
- Continuous valued MRFs
  - Inference using Harmonic solution
  - Structure estimation through l-1 penalized MLE
- Binary valued MRFs (Ising model)
  - Inference
    - Gaussian approximation – Random walk interpretation
    - MCMC
  - Structure estimation
    - Pseudo MLE
    - Logistic regression
- Conclusions

Directed Acyclical GMs (Bayesian networks)

- Ordered Markov property : $x_s \perp \mathbf{x}_{\mathrm{pred}(s)\setminus \mathrm{pa}(s)} | \mathbf{x}_{\mathrm{pa}(s)}$
  - Complete independence: Markov "Blanket" (Parents+children+co-parents)
- Joint pdf modeled as product of conditionals: $p(\mathbf{x}_{1:V}|G) = \prod_{t=1}^{V} p(x_t|\mathbf{x}_{\mathrm{pa}(t)})$
- Examples

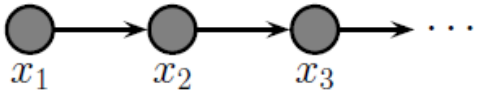
Naïve Bayes



$$p(y, \mathbf{x}) = p(y) \prod_{j=1}^{D} p(x_j|y)$$

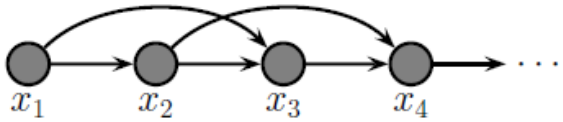
Markov chain

$$p(x_{1:T}) = p(x_1) \prod_{t=2}^{T} p(x_t|x_{t-1})$$

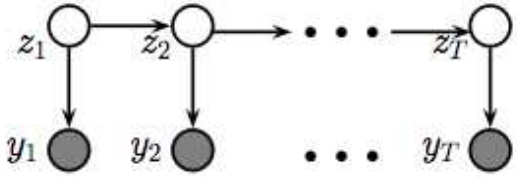


2nd order Markov chain

$$p(x_{1:T}) = p(x_1, x_2) \prod_{t=3}^{T} p(x_t|x_{t-1}, x_{t-2})$$

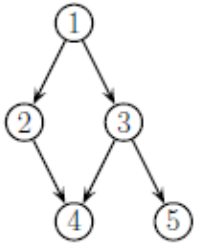


Hidden Markov model



$$\begin{aligned} p(y_{1:T}, z_{1:T}) &= p(\mathbf{z}_{1:T}) p(\mathbf{y}_{1:T}|\mathbf{z}_{1:T}) \\ &= \left[p(z_1) \prod_{t=2}^{T} p(z_t|z_{t-1}) \right] \left[\prod_{t=1}^{T} p(y_t|z_y) \right] \end{aligned}$$

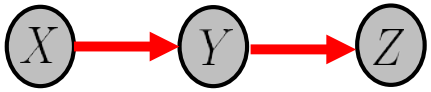
Arbitrary



$$\begin{aligned} p(\mathbf{x}_{1:5}) &= p(x_1)p(x_2|x_1)p(x_3|x_1, \cancel{x_2})p(x_4|\cancel{x_1}, x_2, x_3)p(x_5|\cancel{x_1}, \cancel{x_2}, x_3, \cancel{x_4}) \\ &= p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2, x_3)p(x_5|x_3) \end{aligned}$$

Basic building blocks of Bayesian nets

- The chain structure

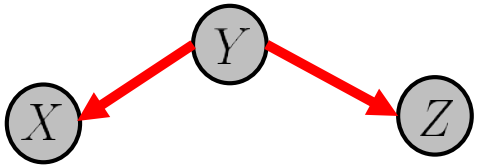


$$p(x,y,z) = p(x)p(y|x)p(z|y)$$

$$p(x,z|y) = \frac{p(x)p(y|x)p(z|y)}{p(y)} = \frac{p(x,y)p(z|y)}{p(y)} = p(x|y)p(z|y)$$

$$X \perp Z|Y$$

- The tent structure

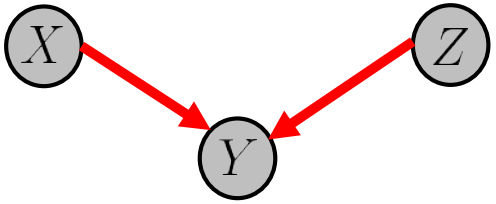


$$p(x,y,z) = p(y)p(x|y)p(z|y)$$

$$p(x,z|y) = \frac{p(x,y,z)}{p(y)} = \frac{p(y)p(x|y)p(z|y)}{p(y)} = p(x|y)p(z|y)$$

$$X \perp Z|Y$$

- The V structure



$$p(x,y,z) = p(x)p(z)p(y|x,z)$$

$$p(x,z|y) = \frac{p(x)p(z)p(y|x,z)}{p(y)}$$

$$X \not\perp Z|Y$$

$$p(x,z) = p(x)p(z)$$

$$X \perp Y$$

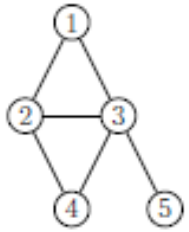
Berkson's Paradox ("explaining away")

Undirected GMs (Markov random fields)

- More natural in some domains (e.g. special statistics, relational data)
  - Simple rule: Nodes not connected w. edge are conditionally independent
- Joint pdf parametrized and modeled as product of factors(not conditionals)
  - Each factor $\psi_c(\mathbf{x}_c;\theta_c)$ or potential corresponds to a maximal clique c
- Hamersley-Clifford theorem
  - $p(\mathbf{x})$ satisfies the CI properties of an undirected graph *iff*

$$p(\mathbf{x};\theta) = \frac{1}{Z(\theta)} \prod_{c\in\mathcal{C}} \psi_c(\mathbf{x}_c;\theta_c) \quad \text{where} \quad Z(\theta) := \sum_{\mathbf{x}} \prod_{c\in\mathcal{C}} \psi_c(\mathbf{x}_c;\theta_c)$$

- Example



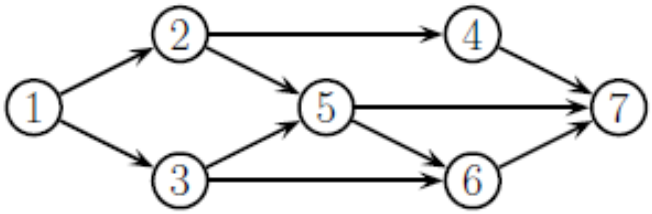
$$p(\mathbf{y}|\theta) = \frac{1}{Z(\theta)} \psi_{123}(y_1,y_2,y_3)\psi_{234}(y_2,y_3,y_4)\psi_{35}(y_3,y_5)$$

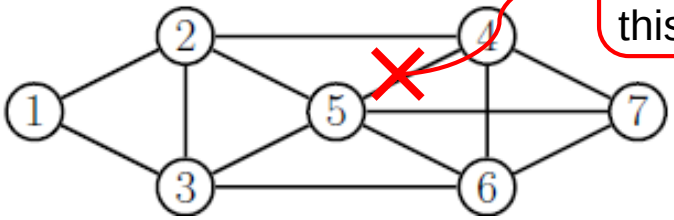
Partition Function: $Z = \sum_{\mathbf{y}} \psi_{123}(y_1,y_2,y_3)\psi_{234}(y_2,y_3,y_4)\psi_{35}(y_3,y_5)$

Generally NP-hard to compute

Equivalence of DGMs and UGMs

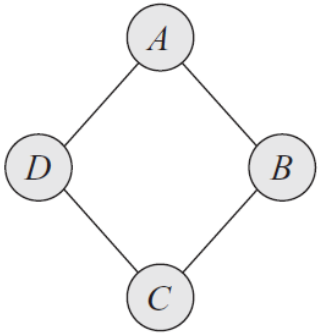
- Moralization: Transition from directed to undirected GM
  - Drop directionality and connect “unmarried” parents
  - Information may be lost during transition (see example)



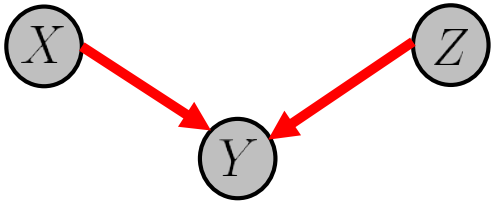


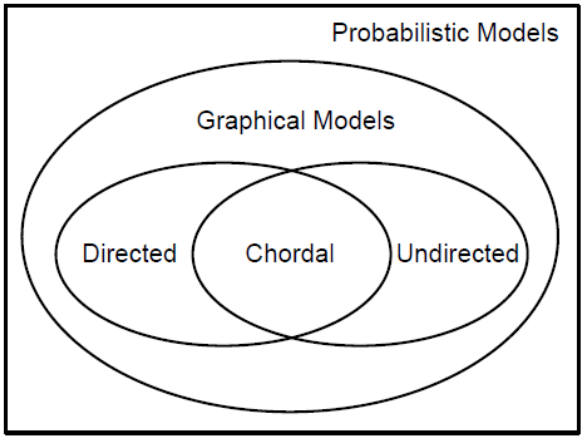
$4 \perp 5 | 2$ lost due to this edge

Cannot be represented by DAGs



Cannot be represented by UGMs





MRFs with energy functions

- Clique potentials usually represented using an "energy" function $E_c(\mathbf{x}_c)$

$$\psi_c(\mathbf{x}_c;\theta_c) = \exp\left(-E(\mathbf{x}_c;\theta_c)\right)$$

- Joint (Gibbs distribution)

$$p(\mathbf{x};\theta) = \frac{1}{Z(\theta)} \exp\left(-\sum_c E(\mathbf{x}_c;\theta_c)\right)$$

- High probability states correspond to low energy configurations
- Any MRF can be decomposed to pairwise potentials (and energy functions)

$$p(\mathbf{x};\theta) = \frac{1}{Z(\theta)} \exp\left(-\sum_{i,j} \phi(x_i,x_j)\theta_{ij}\right)$$

- MRF is associative if $\phi(x_i,x_j)$ measures difference btw x_i and x_j, and $\theta \geq 0$
  - Gaussian MRF: $\phi(x_i,x_j) = (x_i - x_j)^2$
  - Ising (binary +1,-1) model: $\phi(x_i,x_j) = x_i x_j$

Gaussian MRFs

- Joint Gaussian fully parametrized by covariance and mean

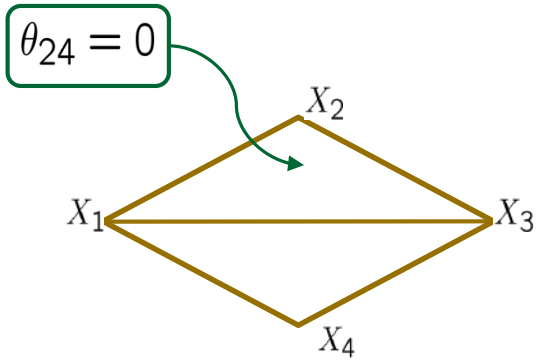
$$p(\boldsymbol{x};\boldsymbol{\Sigma}) = \frac{1}{2^{N/2}\det(\boldsymbol{\Sigma})^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

- GMRF structure given by precision matrix (inv. Cov.)
  - Also viewed as the Laplacian of the graph

$$\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1} = ((\theta_{ij}))$$

$$p(\boldsymbol{x};\boldsymbol{\Theta}) = \frac{\det(\boldsymbol{\Theta})^{1/2}}{2^{N/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Theta}(\boldsymbol{x}-\boldsymbol{\mu})}$$

$$x_i \perp x_j \,|\, \mathbf{x}_{\mathcal{V}-i,j} \iff \theta_{i,j} = 0$$



- Assume for simplicity (and wlog) that $\mu = \mathbf{0}$

- Inference: Given known $\boldsymbol{\Theta}$ and observed $\mathbf{x}_s$, find $p(\mathbf{x}_{s^c}|\mathbf{x}_s)$

Inference via Harmonic solution

- Negative log-likelihood of joint

$$-\log p(\mathbf{x}_{s^c}, \mathbf{x}_s) \propto \begin{bmatrix} \mathbf{x}_{s^c}^T & \mathbf{x}_s^T \end{bmatrix} \begin{bmatrix} \boldsymbol{\Theta}_{s^c,s^c} & \boldsymbol{\Theta}_{s^c,s} \\ \boldsymbol{\Theta}_{s,s^c} & \boldsymbol{\Theta}_{s,s} \end{bmatrix} \begin{bmatrix} \mathbf{x}_{s^c} \\ \mathbf{x}_s \end{bmatrix}$$

$$= \mathbf{x}_{s^c}^T \boldsymbol{\Theta}_{s^c,s^c} \mathbf{x}_{s^c} + 2\mathbf{x}_s^T \boldsymbol{\Theta}_{s,s^c} \mathbf{x}_{s^c} + \mathbf{x}_s^T \boldsymbol{\Theta}_{s,s} \mathbf{x}_s + \mathbf{x}_{s^c}^T \boldsymbol{\Theta}_{s^c,s} \boldsymbol{\Theta}_{s,s}^{-1} \boldsymbol{\Theta}_{s,s^c} \mathbf{x}_{s^c} - \mathbf{x}_{s^c}^T \boldsymbol{\Theta}_{s^c,s} \boldsymbol{\Theta}_{s,s}^{-1} \boldsymbol{\Theta}_{s,s^c} \mathbf{x}_{s^c}$$

$$= \mathbf{x}_{s^c}^T \boldsymbol{\Theta}_{s^c,s^c} \mathbf{x}_{s^c} + 2\mathbf{x}_s^T \boldsymbol{\Theta}_{s,s^c} \mathbf{x}_{s^c} + \mathbf{x}_s^T \boldsymbol{\Theta}_{s,s^c} \boldsymbol{\Theta}_{s^c,s^c}^{-1} \boldsymbol{\Theta}_{s^c,s} \mathbf{x}_s \rightarrow -\log p(\mathbf{x}_{s^c}|\mathbf{x}_s)$$

$$+\mathbf{x}_s^T \left(\boldsymbol{\Theta}_{s,s} - \boldsymbol{\Theta}_{s,s^c} \boldsymbol{\Theta}_{s^c,s^c}^{-1} \boldsymbol{\Theta}_{s^c,s} \right) \mathbf{x}_s \rightarrow -\log p(\mathbf{x}_s)$$

- Finally

$$-\log p(\mathbf{x}_{s^c}|\mathbf{x}_s) = \left(\mathbf{x}_{s^c} + \boldsymbol{\Theta}_{s^c,s^c}^{-1} \boldsymbol{\Theta}_{s^c,s} \mathbf{x}_s \right)^T \boldsymbol{\Theta}_{s^c,s^c} \left(\mathbf{x}_{s^c} + \boldsymbol{\Theta}_{s^c,s^c}^{-1} \boldsymbol{\Theta}_{s^c,s} \mathbf{x}_s \right)$$

$$p(\mathbf{x}_{s^c}|\mathbf{x}_s) \sim \mathcal{N}\left(-\boldsymbol{\Theta}_{s^c,s^c}^{-1} \boldsymbol{\Theta}_{s^c,s} \mathbf{x}_s, \boldsymbol{\Theta}_{s^c,s^c} \right) \rightarrow \text{“Harmonic”}$$

- Conditional mean of $\mathbf{x}_{s^c}$ contains all information from observed $\mathbf{x}_s$

GMRF structure estimation via maximum likelihood

- Given $\{\boldsymbol{x}^{(t)}\}_{t=1}^{T}$, goal is to estimate $\boldsymbol{\Theta}$ and μ

$$p(\boldsymbol{x};\boldsymbol{\Theta}) = \frac{\det(\boldsymbol{\Theta})^{1/2}}{2^{N/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Theta}(\boldsymbol{x}-\boldsymbol{\mu})}$$

$$\Longrightarrow \quad \log p(\boldsymbol{x};\boldsymbol{\Theta}) = \frac{1}{2}\log\det(\boldsymbol{\Theta}) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Theta}(\boldsymbol{x}-\boldsymbol{\mu}) + c$$

- Log-likelihood

$$\mathcal{L}(\boldsymbol{\Theta}) = \max_{\mu} \frac{1}{T}\sum_{n=1}^{T}\left(\log\det(\boldsymbol{\Theta}) - \left(\boldsymbol{x}^{(t)}-\boldsymbol{\mu}\right)^T\boldsymbol{\Theta}\left(\boldsymbol{x}^{(t)}-\boldsymbol{\mu}\right)\right)$$

$$= \log\det(\boldsymbol{\Theta}) - \frac{1}{T}\sum_{t=1}^{N}\left(\boldsymbol{x}^{(t)}-\hat{\boldsymbol{\mu}}\right)^T\boldsymbol{\Theta}\left(\boldsymbol{x}^{(t)}-\hat{\boldsymbol{\mu}}\right)$$

$$= \log\det(\boldsymbol{\Theta}) - \operatorname{trace}\left(\underbrace{\frac{1}{T}\sum_{t=1}^{T}\left(\boldsymbol{x}^{(t)}-\hat{\boldsymbol{\mu}}\right)\left(\boldsymbol{x}^{(t)}-\hat{\boldsymbol{\mu}}\right)^T}_{\boldsymbol{S}}\boldsymbol{\Theta}\right)$$

$$\hat{\boldsymbol{\Theta}}_{ML} = \arg\max_{\boldsymbol{\Theta}\succ\mathbf{0}} \log\det(\boldsymbol{\Theta}) - \operatorname{trace}(\boldsymbol{S}\boldsymbol{\Theta})$$

ℓ_1-penalized MLE of Θ

$$\hat{\Theta}_{ML} = \arg\max_{\Theta \succ 0} \log\det(\Theta) - \text{trace}(S\Theta)$$

- Closed-form solution: $\Theta^{-1} - S = 0 \Rightarrow \hat{\Theta}_{ML} = S^{-1}$
- $\hat{\Theta}_{ML}$ generally is full matrix
- Idea: Add ℓ_1 constrain on to enforce (sparse) graph structure

$$\hat{\Theta}_{\ell_1} = \arg\min_{\Theta \succ 0} -\log\det(\Theta) + \text{trace}(S\Theta) + \lambda ||\Theta||_1$$

$$||\Theta||_1 = \sum_{i=1}^{P} \sum_{j=1}^{P} |\theta_{ij}|$$

- Problem is convex and for $W = \Theta^{-1}$ is equivalent to

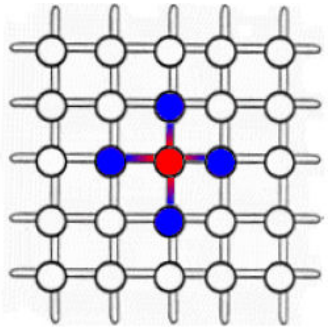
$$\hat{W} = \arg\max_{W \succ 0} \log\det(W) \quad \text{s.to } ||W - S||_\infty \leq \lambda$$

Solvable via Graphical Lasso

O. Banerjee, L. El Ghaoui, and A. d'Aspremont, "Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data," *J. Machine Learning Research*, vol. 9, pp. 485-516, June 2008.

Binary random variables

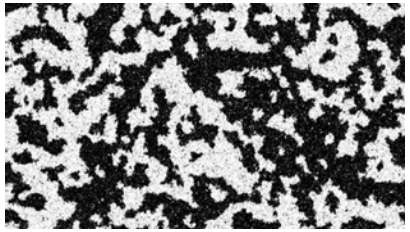
Ising model



- Ising model for $\boldsymbol{x} \in \{-1, 1\}^N$ or $\boldsymbol{x} \in \{0, 1\}^N$

$$P(\boldsymbol{x}; \boldsymbol{\Theta}) = \exp\left(\boldsymbol{x}^T \boldsymbol{\Theta} \boldsymbol{x} - Z(\boldsymbol{\Theta})\right) \quad \theta_{ii} = 0$$

Log partition function: $Z(\boldsymbol{\Theta}) = \log \left(\sum_{\boldsymbol{x} \in \{-1,1\}^N} \exp\left(\boldsymbol{x}^T \boldsymbol{\Theta} \boldsymbol{x}\right) \right)$



- Estimation: l-1 penalized maximum likelihood for $\boldsymbol{\Theta}$

$$\hat{\boldsymbol{\Theta}}_{ML} = \arg\min_{\boldsymbol{\Theta}} Z(\boldsymbol{\Theta}) - \text{trace}(\boldsymbol{\Theta} \boldsymbol{S}) + \lambda ||\boldsymbol{\Theta}||_1$$

$$\boldsymbol{S} = \frac{1}{N} \sum_{t=1}^{T} \boldsymbol{x}^{(t)} \boldsymbol{x}^{(t)T}$$

- Problem: $Z(\boldsymbol{\Theta})$ combinatorialy complex to compute
- Two alternatives: $Z(\boldsymbol{\Theta})$ is upper-bounded or avoided
- Similar problem for inference: $p(\mathbf{x}_{s^c} | \mathbf{x}_s)$ can only be approximated

The role of Θ in Ising model

Claim: $\theta_{ij} = 0 \iff x_i \perp x_j | \{X_k,\ k \neq i,j\}$

- Proof: consider $\boldsymbol{x} = (x_i, \boldsymbol{x}_{-i})$ and $\Theta = \begin{bmatrix} 0 & \boldsymbol{\theta}_i^T \\ \boldsymbol{\theta}_i & \Theta_{-i} \end{bmatrix}$

$$P(x_i = a|\boldsymbol{x}_{-i}) = \frac{P(x_i=a,\boldsymbol{x}_{-i})}{P(x_i=a,\boldsymbol{x}_{-i})+P(x_i=-a,\boldsymbol{x}_{-i})} \qquad \forall a \in \{-1,1\}$$

$$= \frac{1}{1+P(x_i=-a,\boldsymbol{x}_{-i})/P(x_i=a,\boldsymbol{x}_{-i})} = \gamma$$

- Use the Ising model

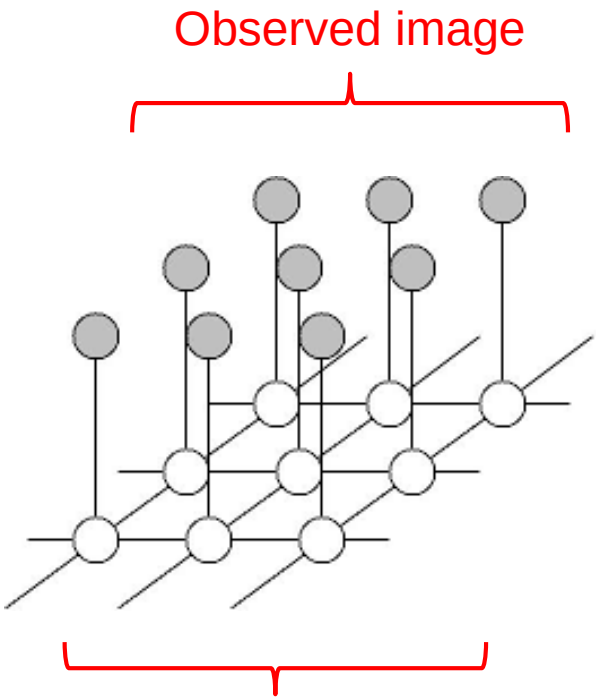
$$\gamma = \frac{\exp(-a\boldsymbol{\theta}_i^T\boldsymbol{x}_{-i} + \frac{1}{2}\boldsymbol{x}_{-i}^T\Theta_{-i}\boldsymbol{x}_{-i} - Z(\Theta))}{\exp(a\boldsymbol{\theta}_i^T\boldsymbol{x}_{-i} + \frac{1}{2}\boldsymbol{x}_{-i}^T\Theta_{-i}\boldsymbol{x}_{-i} - Z(\Theta))} = \exp\left(-2a\boldsymbol{\theta}_i^T\boldsymbol{x}_{-i}\right)$$

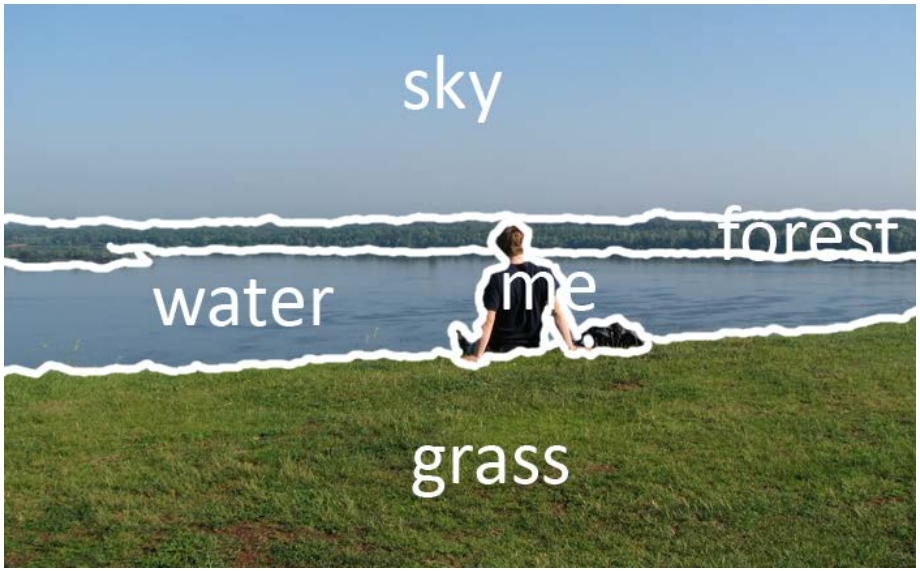
- Plug γ in the expression above

$$P(x_i = a|\boldsymbol{x}_{-i}) = \frac{1}{1+\exp(-2a\boldsymbol{\theta}_i^T\boldsymbol{x}_{-i})} = \frac{1}{1+\exp(-2a\sum_{j\neq i}\theta_{ij}x_j)}$$

Example: Image segmentation

- Use 2-D HMM (Ising as hidden layer) to infer “meaning” of image pixels





Inference via Gaussian field approximation

- Exact inference NP-hard
- Use surrogate continuous-values Gaussian random field: $\tilde{\mathbf{x}}_{s^c} \sim \mathcal{N}(\mu_{s^c}, \mathbf{\Theta}_{s^c,s^c})$
  - Compute exact Harmonic solution: $\mu_{s^c} = -\mathbf{\Theta}_{s^c,s^c}^{-1}\mathbf{\Theta}_{s^c,s}\mathbf{x}_s$
- Predictor of unknown labels via GMRF mean: $\hat{x}_i = \begin{cases} 1 & \mu_i > 1/2 \\ 0 & \text{else} \end{cases} \quad \forall i \in \mathcal{S}^c$
- Approximation of marginal posteriors: $p\left(x_i = 1 | \mathbf{x}_s\right) \approx \mathbb{E}\left[\tilde{x}_i\right] = \mu_i \quad \forall i \in \mathcal{S}^c$
- Random walk interpretation
  - Imagine particle performing a random walk on (unobserved) graph
  - Let normalized Laplacian $\mathbf{P} = \mathbf{D}^{-1}\mathbf{\Theta}_{s^c,s^c}$ be transition probability matrix
  - Observed variables act as sink nodes where the walk ends
  - Starting from node i, probability that walk ends in +1 node is μ_i

Inference via MCMC

- Collect samples $\{\mathbf{x}_{s^c}^{(t)}\}_{t=1}^{T}$ from MC with $p(\mathbf{x}_{s^c}|\mathbf{x}_s)$ as stationary distribution
- Gibbs sampler: One variable (node) sampled at every round t (the rest are fixed)
  - Exploits (sparse) conditional dependence structure of MRF
  - Observed nodes used as (fixed) boundary conditions

$$x_i^{(t)} \sim \mathrm{Ber}(p_i^{(t)})$$

$$p_i^{(t)} = \begin{cases} \left(1+\exp(-2\sum_{j\in N(i)}\theta_{ij}x_j^{(t-1)})\right)^{-1}, & i \in \mathcal{S}^c \\ \mathbb{I}_{\{x_i=1\}}, & i \in \mathcal{S} \end{cases}$$

$$p(x_i = 1|\mathbf{x}_{s^c}) \approx \frac{1}{T}\sum_{t=1}^{T}\mathbb{I}_{\{x_i^{(t)}=1\}}\ , \ \forall i \in \mathcal{S}$$

$$\begin{array}{cccc} x_1^{(1)} & x_2^{(1)} & \dots & x_N^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \dots & x_N^{(2)} \\ x_1^{(3)} & x_2^{(3)} & \dots & \\ \vdots & & & \end{array}$$

- Experiments indicate Gibbs smpl offers better inference in rect. Ising models
- More sophisticated MCMC methods achieve faster mixing (e.g. Wolfs algorithm)

Towards estimation: Bounding the partition function

- Goal: Find $\bar{Z}(\Theta) \geq Z(\Theta)$ computable with polynomial complexity

$$Z(\Theta) = \log \left(\sum_{x \in \{0,1\}^N} \exp(x^T \Theta x) \right)$$

- Consider partition $\mathcal{B}_k \subset \{0,1\}^N$ such that $x \in \mathcal{B}_k \iff ||x||_0 = k$

$$Z(\Theta) = \log \left(\sum_{k=0}^{N} \sum_{x \in \mathcal{B}_k} \exp(x^T \Theta x) \right) \leq \log \left(\sum_{k=0}^{N} \sum_{x \in \mathcal{B}_k} \exp \left(\max_{x \in \mathcal{B}_k} x^T \Theta x \right) \right)$$

$$= \log \left(\sum_{k=0}^{N} \binom{N}{k} \exp \left(\phi(\Theta) \right) \right)$$

- Computing $\phi(\Theta) = \max_{x \in \mathcal{B}_k} x^T \Theta x$ is still hard

L. El Ghaoui, A. Gueye. “A Convex Upper Bound on the Log-Partition Function for Binary Graphical Models,” *Journal of Machine Learning Research*, vol. 9, pp. 485–516, Mar. 2008.

Relaxation of the bound

- Relax $\phi(\Theta) = \max_{x \in \mathcal{B}_k} x^T \Theta x = \max_{x \in \mathcal{B}_k} \text{trace}(\Theta x x^T)$
- Add redundant constrains

$$\phi(\Theta) = \max_{X \in \mathcal{S}_+, x \in \mathcal{B}_k} \text{trace}(\Theta X)$$
$$\text{s. to } X \succeq xx^T,\ x^T x = k,\ 1^T mX1 = k^2,\ \text{diag}(X) = 1,\ \text{rank}(X) = 1$$

- Relax

$$\psi(\Theta) = \max_{X \in \mathcal{S}_+, x} \text{trace}(\Theta X)$$
$$\text{s. to } X \succeq xx^T,\ x^T x = k,\ 1^T X1 = k^2,\ \text{diag}(X) = x$$

- Upper-bound $\psi(\Theta) \geq \phi(\Theta)$

$$\Rightarrow Z(\Theta) \leq \log \left(\sum_{k=0}^{N} \binom{N}{k} \exp\left(\phi(\Theta)\right) \right) \leq \log \left(\sum_{k=0}^{N} \binom{N}{k} \exp\left(\psi(\Theta)\right) \right) := \bar{Z}(\Theta)$$

- Claim: bound quality $0 \leq \bar{Z}(\Theta) - Z(\Theta) \leq \min_{\mu,\lambda} ||\Theta - \mu I - \lambda 11^T||_1$

Pseudo Maximum Likelihood

- Want to solve: $\hat{\Theta}_{PML} = \arg\min_{\Theta} \bar{Z}(\Theta) - \text{trace}(\boldsymbol{S}\Theta) + \lambda||\Theta||_1$

$$\hat{\Theta}_{PML} = \arg\min_{\Theta} \log\left(\sum_{k=0}^{N} \binom{N}{k} \exp\left(\psi(\Theta)\right)\right) - \text{trace}(\boldsymbol{S}\Theta) + \lambda||\Theta||_1$$

- Dual

$$\psi(\Theta) = \min_{t,\mu,\boldsymbol{\nu},\lambda} t + \mu k + \lambda k^2 \quad \text{s. to} \begin{pmatrix} \text{Diag}(\boldsymbol{\nu}) + \mu\boldsymbol{I} + \lambda\mathbf{1}\mathbf{1}^T - \Theta & \frac{1}{2}\boldsymbol{\nu} \\ \frac{1}{2}\boldsymbol{\nu}^T & t \end{pmatrix} \succeq 0$$

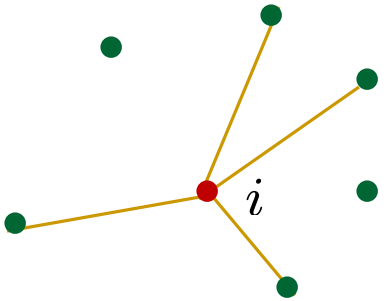
- Substituting dual $\psi(\Theta)$ above

$$\hat{\Theta}_{PML} = \arg\min_{\Theta,t,\mu,\boldsymbol{\nu},\lambda} \log\left(\sum_{k=0}^{N} \binom{N}{k} \exp\left(t + \mu k + \lambda k^2\right)\right) - \text{trace}(\boldsymbol{S}\Theta) + \lambda||\Theta||_1$$

$$\text{s. to} \begin{pmatrix} \text{Diag}(\boldsymbol{\nu}) + \mu\boldsymbol{I} + \lambda\mathbf{1}\mathbf{1}^T - \Theta & \frac{1}{2}\boldsymbol{\nu} \\ \frac{1}{2}\boldsymbol{\nu}^T & t \end{pmatrix} \succeq 0$$

Logistic regression for Θ

- Goal: Estimate Θ while avoiding computation of $Z(\Theta)$



- Idea: consider node i and its connections
  - Separate $\boldsymbol{x} = (x_i, \boldsymbol{x}_{-i})$
  - Use $\boldsymbol{x}_{-i}$ as input and x_i as output
  - Logistic regression ⟹ parametric estimation of $\log P(x_i|\boldsymbol{x}_{-i})$
  - Estimate $\boldsymbol{\theta}_i$ as a byproduct

- Problem statement: re-write problem bellow for the Ising model

$$\hat{\boldsymbol{\theta}}_i = \arg\min_{\boldsymbol{\theta}_i} - \sum_{t=1}^{T} \log P_{\boldsymbol{\theta}_i}\left(x_i = x_i^{(t)} | \boldsymbol{x}_{-i} = \boldsymbol{x}_{-i}^{(t)}\right) + \lambda ||\boldsymbol{\theta}_i||_1$$

P. Ravikumar, M. J. Wainwright and J. Lafferty. *High-dimensional Ising model selection using ℓ_1-regularized logistic regression.* To appear in the Annals of Statistics. Available at http://www.eecs.berkeley.edu

Estimation of Θ

- We have: $$P(x_i|\boldsymbol{x}_{-i}) = \frac{1}{1+\exp(-2x_i\boldsymbol{\theta}_i^T\boldsymbol{x}_{-i})}$$
$$= \frac{\exp(x_i\boldsymbol{\theta}_i^T\boldsymbol{x}_{-i})}{\exp(x_i\boldsymbol{\theta}_i^T\boldsymbol{x}_{-i}) + \exp(-x_i\boldsymbol{\theta}_i^T\boldsymbol{x}_{-i})}$$

- Taking the logarithm

$$\Rightarrow \log P(x_i|\boldsymbol{x}_{-i}) = x_i\boldsymbol{\theta}_i^T\boldsymbol{x}_{-i} - \log\left(\exp(x_i\boldsymbol{\theta}_i^T\boldsymbol{x}_{-i}) + \exp(-x_i\boldsymbol{\theta}_i^T\boldsymbol{x}_{-i})\right)$$

- Substituting the log-likelihood

$$\hat{\boldsymbol{\theta}}_i = \arg\min_{\boldsymbol{\theta}_i} -\sum_{t=1}^{T} \log P\left(x_i = x_i^{(t)} | x_{-i} = x_{-i}^{(t)}\right) + \lambda||\boldsymbol{\theta}_i||_1$$
$$= \arg\min_{\boldsymbol{\theta}_i} \sum_{t=1}^{T} \left[\log\left(\exp(x_i^{(t)}\boldsymbol{\theta}_i^T\boldsymbol{x}_{-i}^{(t)}) + \exp(-x_i^{(t)}\boldsymbol{\theta}_i^T\boldsymbol{x}_{-i}^{(t)})\right) - x_i^{(t)}\boldsymbol{\theta}_i^T\boldsymbol{x}_{-i}^{(t)}\right] + \lambda||\boldsymbol{\theta}_i||_1$$

- Convex problem

Conclusions

- Graphical models
  - Modeling pdfs using conditional dependencies
  - Undirected models (MRFs) naturally modeled by graphs
  - Inference in closed form for Gaussian MRFs
  - Estimation of GMRFs as Laplacian fitting problem
  - Inference and estimation approximations for binary MRFs (Ising model)

- Possible research directions
  - Active sampling on binary MRFs using MCMC
  - Active sampling for MRF structure estimation