ML for UGM:

data $D = \{\vec{x}^1, \ldots, \vec{x}^N\}$

I am using $\rightarrow$ on the top to indicate it is a multivariate variable (i.e. $\dim(\vec{X}) > 1$). I use superindex to represent observations. $\vec{x}^1, \ldots, \vec{x}^N \overset{i.i.d}{\sim} P(\vec{x}; \theta)$

$$\max_\theta \log P(D; \theta) = \max_\theta \sum_{n=1}^{N} \log p(\vec{x}^n; \theta) \quad (*)$$

θ: all Params

$$P(\vec{x}; \theta) = \underbrace{\frac{1}{Z(\theta)}}_{\text{normalizer}} \prod_c \underbrace{\phi_c(x_c; \theta)}_{\text{factors}}$$

$$(*) \Rightarrow \sum_{n=1}^{N} \sum_c \log \phi_c(x_c^n; \theta) - N \log Z(\theta)$$

the trouble maker!!!

why?

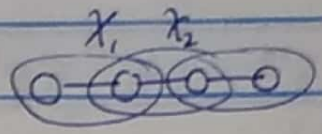
$$L(\theta) = \sum_{n=1}^{N} \sum_c \log \phi_c(x_c^n; \theta) - N \log \left(\sum_{x} \prod_{c'} \phi_{c'}(x_{c'}; \theta) \right)$$

x: not dependent on data

Page 2

Derivative of ML (w.r.t. θ_c):

$$\nabla_{\theta_c} L(\theta) =$$



$$\sum_{n=1}^{N} \frac{\partial}{\partial \theta_c} \log \phi(x_c; \theta_c) - N \frac{\partial}{\partial \theta_c} \log Z(\theta)$$

$$\frac{\partial}{\partial \theta_c} Z(\theta) = \frac{\partial}{\partial \theta_c} \sum_y \prod_{c'} \phi_{c'}(y_{c'}; \theta_{c'})$$

$$= \sum_y \frac{\partial \phi_c(y_c; \theta_c)}{\partial \theta_c} \prod_{c' \neq c} \phi_{c'}(y_{c'}; \theta_{c'})$$

Mult/divide $\phi_c(y_c; \theta_c)$

$$= \sum_y \frac{\partial \phi_c(y_c; \theta_c)}{\partial \theta_c} \cdot \frac{1}{\phi_c(y_c; \theta_c)} \prod_c \phi_{c'}(y_{c'}; \theta_{c'})$$

Mult/divide $Z(\theta)$

$$= Z(\theta) \sum_y \underbrace{\frac{\partial \theta_c(y_c; \theta_c)}{\partial \theta_c} \cdot \frac{1}{\phi_c(y_c; \theta_c)}}_{\frac{\partial}{\partial \theta_c} \log \phi_c(y_c; \theta_c)} \underbrace{\frac{\prod_c \phi_c(y_c; \theta_c)}{Z(\theta)}}_{p(y; \theta)}$$

$$= Z(\theta) \underbrace{\mathbb{E}}_{y_c \sim P(y_c)} \left[\frac{\partial}{\partial \theta_c} \log \phi_c(y_c; \theta_c) \right]$$

marginal

Substitude it back:

$$\frac{\partial}{\partial \theta_c} L(\theta) = \sum_{n=1}^{N} \frac{\partial}{\partial \theta_c} \log \phi_c(x_c^n; \theta_c) - N \mathbb{E}_{y_c \sim P(y_c)} \left[\frac{\partial}{\partial \theta_c} \log \phi(y_c; \theta_c) \right]$$

empirical mean — actual mean

at the optimum ($\nabla_{\theta_c} L(\theta) = 0$), the empirical and actual moments match!

For tabular clique Potential:

$$L(\phi) = \sum_n \sum_c \sum_{y_c} \mathbb{I}(y_c = x_c^n) \log \phi_c(y_c) - N \log Z(\phi)$$

$$\nabla_{\phi_c} L(\phi) = \sum_n \mathbb{I}(y_c = x_c^n) \frac{1}{\phi_c(y_c)} - N \nabla_{\phi_c} \log Z(\theta)$$

$$\nabla_{\phi_c} \log Z(\theta) = \nabla_{\phi_c} \log \left(\sum_{y_{c'}} \prod_{c'} \phi_{c'}(y_{c'}) \right)$$

$$= \frac{\sum_{y_{c'}} \prod_{c' \neq c} \phi_{c'}(y_{c'})}{\sum_{y_{c'}} \prod_{c'} \phi_{c'}(y_{c'})} = \frac{\sum_{y_{c'}} \prod_{c'} \phi_{c'}(y_{c'})}{Z(\theta)\, \phi_c(y_c)}$$

$$= \frac{P(y_c)}{\phi_c(y_c)}$$

$$\nabla_{\phi_c} L(\phi) = \sum_n \mathbb{I}(y_c = x_c^n) \frac{1}{\phi_c(y_c)} - N \frac{P(y_c)}{\phi_c(y_c)}$$

Let's introduce this notation for empirical mean:

$$\varepsilon(x_c^n) = \frac{1}{N} \sum \mathbb{I}(y_c = x_c^n)$$

at the optimal param $\nabla_{\phi_c} L(\phi)$:

$$\varepsilon(x_c^n) = P(y_c)$$

⊠

$$\log P(x;\theta) = \sum_c \theta_c^T f(x_c) - \log Z(\theta)$$

$$\nabla_{\theta_c}(\quad) = f(x_c) - \frac{\nabla_{\theta_c} Z(\theta)}{Z(\theta)}$$

$$Z(\theta) = e^{A(\theta)} \qquad \nabla_{\theta_c} Z(\theta) = \frac{\partial A(\theta)}{\partial \theta_c} e^{A(\theta)}$$

$$= \mathbb{E}[f_c(x_c)] Z(\theta)$$

log-likelihood $L(\theta) = \log P(D;\theta) = \frac{1}{N_c} \sum_{n=1}^{N} \theta_c^T f(x_c^n) - \log Z(\theta)$

$$\nabla_\theta L(\theta) = \frac{1}{N} \sum^{N} f(x_c^n) - \mathbb{E}[f_c(x_c)]$$

Page 5

$$= \mathbb{E}_{x_c \sim \tilde{P}(\cdot;\theta_c)}[f(x_c)] - \mathbb{E}_{x_c \sim P(\cdot;\theta)}[f(x_c)]$$

Empirical