$$\ell(\theta) = \prod_{i=1}^{m} P_\theta(x^i) \qquad D = x^1, x^2, \cdots, x^m$$

$$\max_\theta \Pr(\underbrace{x^1, x^2, -, x^m}_{D} | \theta)$$

$$\max_\theta \Pr(D|\theta)$$

Bayesian Learning

prior knowledge

$$\max_\theta \Pr(\theta|D) = \frac{\Pr(D|\theta)\, P(\theta)}{P(D)}$$

$$\theta^* = \arg\max_\theta \frac{\Pr(D|\theta)\, P(\theta)}{P(D)} = \arg\max_\theta P(D|\theta)\, P(\theta)$$

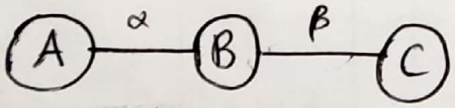
$$= \arg\max_\theta \ell(\theta)\, P(\theta)$$

$$= \arg\max_\theta \left[\prod_{i=1}^{m} P_\theta(x^i)\right] P(\theta)$$

$$\log \tilde{P}(\theta|D) = \sum_{i=1}^{m} \log P_\theta(x^i) + \log P(\theta) + \lambda\|\theta\|^2$$

$\lambda\|\theta\|^2$ → regularization

Koller



$\theta = (\alpha, \beta)$

$$P_\theta(A,B,C) = \frac{1}{Z}\phi_\alpha(A,B)\,\phi_\beta(B,C)$$

Data

A^1, B^1, C^1

A^2, B^2, C^2

A^m, B^m, C^m

$$\ell(\theta) = \ell(\alpha,\beta) = \prod_{i=1}^{m} \frac{1}{Z}\phi_\alpha(A^i,B^i)\,\phi_\beta(B^i,C^i)$$

$$ll(\theta) = \log\ell(\theta) = \sum_{i=1}^{m}\left(-\log Z + \log\phi_\alpha(A^i,B^i) + \log\phi_\beta(B^i,C^i)\right)$$

$$= -m\log Z(\theta) + \sum_{i=1}^{m}\log\phi_\alpha(A^i,B^i) + \sum_{i=1}^{m}\log\phi_\beta(B^i,C^i)$$

$$Z = \sum_C\sum_B\sum_A \phi_\alpha(A,B)\,\phi_\beta(B,C) = Z(\alpha,\beta) = Z(\theta)$$

$$ll(\theta) = -m\log\sum_C\sum_B\sum_A\left[\phi_\alpha(A,B)\,\phi_\beta(B,C)\right] + \sum_{i=1}^{m}\log\phi_\alpha(A^i,B^i) + \sum_{i=1}^{m}\log\phi_\beta(B^i,C^i)$$

$P_\theta(X) = \frac{1}{Z(\theta)} \tilde{P}_\theta(X)$ $\quad X \in \mathbb{R}^n$

$P_\theta(X) = \frac{1}{Z(\theta)} \tilde{P}_\theta(X) = \frac{1}{Z(\theta)} e^{F_\theta(X)}$

Data $x^1, x^2, x^3, \ldots, x^m$

$x^i \in \mathbb{R}^n$

$$ll(\theta) = \log \prod_{i=1}^{m} P_\theta(x^i) = \sum_{i=1}^{m} \log \frac{1}{Z(\theta)} \tilde{P}_\theta(x^i) = \sum_{i=1}^{m} \left(-\log Z(\theta) + \log \tilde{P}_\theta(x^i)\right)$$

$\log \equiv \ln$

$$= -m \log Z(\theta) + \sum_{i=1}^{m} \log \tilde{P}_\theta(x^i)$$

$$= -m \log Z(\theta) + \sum_{i=1}^{m} \log \exp(F_\theta(x^i))$$

$$\Rightarrow ll(\theta) = -m \log Z(\theta) + \sum_{i=1}^{m} F_\theta(x^i)$$

$\theta = (\theta_1, \theta_2, \ldots, \theta_p)$ parameters

$\theta^* = \arg\max ll(\theta)$

$$\frac{\partial ll(\theta)}{\partial \theta_k} = -m \frac{\partial}{\partial \theta_k} \log Z(\theta) + \sum_{i=1}^{m} \frac{\partial}{\partial \theta_k} F_\theta(x^i)$$

$$= -m \frac{\frac{\partial}{\partial \theta_k} Z(\theta)}{Z(\theta)} + \sum_{i=1}^{m} \frac{\partial}{\partial \theta_k} F_\theta(x^i)$$

$P_\theta(X) = \frac{1}{Z(\theta)} \exp(F_\theta(X)) \qquad Z(\theta) = \sum_X \exp(F_\theta(X))$

$$\frac{\partial Z(\theta)}{\partial \theta_k} = \sum_X \frac{\partial}{\partial \theta_k} \exp(F_\theta(X)) = \sum_X \left[\frac{\partial}{\partial \theta_k} F_\theta(X)\right] \exp(F_\theta(X))$$

$$\frac{\partial ll(\theta)}{\partial \theta_k} = \frac{-m}{Z(\theta)} \sum_X \left[\frac{\partial}{\partial \theta_k} F_\theta(X)\right] e^{F_\theta(X)} + \sum_{i=1}^{m} \frac{\partial}{\partial \theta_k} F_\theta(x^i)$$

$$= m \left(\frac{1}{m} \sum_{i=1}^{m} \left[\frac{\partial}{\partial \theta_k} F_\theta(x^i)\right] - \sum_X \left[\frac{\partial}{\partial \theta_k} F_\theta(X)\right] \frac{1}{Z(\theta)} e^{F_\theta(X)} \right)$$

$$= m \left(E_D\left\{\frac{\partial}{\partial \theta_k} F_\theta(x^i)\right\} - E_{P_\theta(X)}\left\{\frac{\partial}{\partial \theta_k} F_\theta(X)\right\} \right)$$

positive phase — negative phase

$$P_\theta(X) = \frac{1}{Z(\theta)} e^{F_\theta(X)} \qquad Z(\theta) = \sum_X e^{F_\theta(X)}$$

Data $= D = \{X^1, X^2, \dots, X^m\}$ $\qquad \theta = (\theta_1, \theta_2, \dots, \theta_p)$

$$\frac{\partial}{\partial \theta_k} ll(\theta) = m \left[E_D\left\{ \frac{\partial}{\partial \theta_k} F_\theta(X) \right\} - E_{P_\theta(X)}\left\{ \frac{\partial}{\partial \theta_k} F_\theta(X) \right\} \right]$$

log-linear MRF / No shared parameters

MRF: $F_\theta(X) = \sum_{j=1}^{p} \theta_j f_j(X_{c_j})$

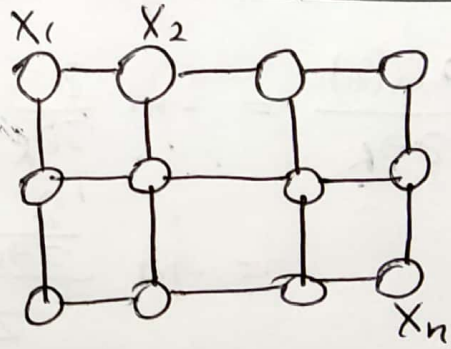
$$\frac{\partial F_\theta(X)}{\partial \theta_k} = f_k(X_{c_j})$$

$$\frac{\partial}{\partial \theta_k} ll(\theta) = m \left[E_D\{ f_k(X_{c_j}) \} - E_{P_\theta(X)}\{ f_k(X_{c_j}) \} \right]$$

Example: Pairwise MRF

No shared parameters

$$P_\theta(X) = \frac{1}{Z} \prod_{i=1}^{m} \phi_i(X_i) \prod_{i,j \in \mathcal{E}} \phi_{ij}(X_i, X_j)$$



$$P_\theta(X) = \frac{1}{Z(\theta)} \exp\left(\sum w_i f_i(X_i) + \sum w_{ij} f_{ij}(X_i, X_j) \right)$$

$\theta = (\{w_i\}, \{w_{ij}\})$

Data:

$X^1 = (X_1^1, X_2^1, \dots, X_n^1)$

$X^2 = (X_1^2, X_2^2, \dots, X_n^2)$

$X^m = (X_1^m, X_2^m, \dots, X_n^m)$

$$\frac{\partial}{\partial w_i} ll(\theta) = m E_D\{ f_i(X_i) \} - m E_{P_\theta}\{ f_i(X_i) \}$$

$$= \sum_{k=1}^{m} f_i(X_i^k) - m \sum_X P_\theta(X_1 - X_m) f_i(X_i)$$

$$= \text{''} \quad - m \sum_{X_1} \sum_{X_2} \cdots \sum_{X_m} P_\theta(X_1, X_2, \dots X_m) f_i(X_i)$$

$$= \text{''} \quad - m \sum_{X_i} \left[\sum_{X_1} \sum_{X_2} \cdots \sum_{X_{i-1}} \sum_{X_{i+1}} \cdots \sum_{X_n} P_\theta(X_1 - X_m) \right] f_i(X_i)$$

$$= \text{''} \quad - m \sum_{X_i} P_\theta(X_i) f(X_i)$$ → marginal distr

$$P_\theta(X) = \frac{1}{Z(\theta)} \exp\Big(\sum_{i=1}^{n} w_i f_i(X_i) + \sum_{(i,j)\in\mathcal{E}} w_{ij} f_{ij}(X_i, X_j) \Big\}$$

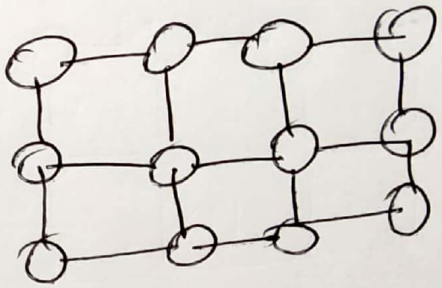
$$\frac{\partial ll(\theta)}{\partial w_i} = \sum_{k=1}^{m} f_i(X_i^k) - m \sum_{X_i} \boxed{P_\theta(X_i)} f_i(X_i)$$

$$\frac{\partial ll(\theta)}{\partial w_{ij}} = \sum_{k=1}^{m} f_{ij}(X_i^k, X_j^k) - m \sum_{X_i} \sum_{X_j} \boxed{P_\theta(X_i, X_j)} f_{ij}(X_i, X_j)$$

To compute the gradient we need the marginal distribution over nodes and edges, i.e. $P_\theta(X_1), P_\theta(X_2), \sim, P_\theta(X_n)$

$P_\theta(X_i, X_j) \quad \forall (i,j) \in \mathcal{E}$

Need inference
exact, approxinate



θ

$$P_\theta(X) = \frac{1}{Z(\theta)} \exp\Big(\sum_{c\in C} \theta_c f(X_c) \Big)$$

$$\frac{\partial ll(\theta)}{\partial \theta_c} = \sum_{k=1}^{m} f(X_c^k) - m \sum_{X_c} P_\theta(X_c) f(X_c)$$

need inferece

init $\theta = \theta_0$

while not converged

- inference $\Rightarrow$ find $P_\theta(X_c)$ for all "c"
- $\theta_t = \theta_{t-1} + \frac{\partial ll(\theta)}{\partial \theta} \longrightarrow \nabla_\theta ll(\theta)$