Probabilistic Graphical Models

Lectures 23,24

Learning MRFs
Contrastive Divergence
Learning CRFs

MRF Log-likelihood

K. N. Toosi University of Technology

$$P_\theta(A, B, C) = 1/Z\ \psi_\alpha(A,B)\ \varphi_\beta(B,C), \qquad \theta = (\alpha,\beta)$$

$$Z(\theta) = Z(\alpha,\beta) = \Sigma_A \Sigma_B \Sigma_C\ \psi_\alpha(A,B)\ \varphi_\beta(B,C)$$

$$ll(\alpha,\beta) = \Sigma_i \log P_\theta(A^i, B^i, C^i)$$

$$= \Sigma_i\ -\log(Z) + \log(\psi_\alpha(A^i,B^i)) + \log(\varphi_\beta(B^i,C^i))$$

$$= -m\ \log(Z) + \Sigma_i \log(\psi_\alpha(A^i,B^i)) + \Sigma_i \log(\varphi_\beta(B^i,C^i))$$

$$= -m\ \log\left(\Sigma_A \Sigma_B \Sigma_C\ \psi_\alpha(A,B)\ \varphi_\beta(B,C)\right)$$

$$+ \Sigma_i \log(\psi_\alpha(A^i,B^i)) + \Sigma_i \log(\varphi_\beta(B^i,C^i))$$

Data:

$$X^1 = (A^1, B^1, C^1)$$

$$X^2 = (A^2, B^2, C^2)$$

$$\vdots$$

$$X^m = (A^m, B^m, C^m)$$

MRF Log-likelihood

K. N. Toosi
University of Technology

$$P_\theta(A, B, C) = 1/Z\ \ \psi_\alpha(A,B)\ \ \varphi_\beta(B,C), \qquad \theta = (\alpha,\beta)$$

$$ll(\alpha,\beta) = -\,m\ \log(Z) + \Sigma_i \log(\psi_\alpha(A^i,B^i)) + \Sigma_i \log(\ \varphi_\beta(B^i,C^i))$$

$$= \underbrace{-\,m\ \log\left(\Sigma_A \Sigma_B \Sigma_C\ \psi_\alpha(A,B)\ \ \varphi_\beta(B,C)\right)}_{f(\alpha,\ \beta)} + \underbrace{\Sigma_i \log(\psi_\alpha(A^i,B^i))}_{g(\alpha,\ data)} + \underbrace{\Sigma_i \log(\ \varphi_\beta(B^i,C^i))}_{h(\beta,\ data)}$$

entangles parameters

MRF log-likelihood - Exponential form

K. N. Toosi University of Technology

CV 23 I

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

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

Data

$X^1, X^2, X^3, \ldots, X^m$

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

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

$\log \equiv \ln$

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

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

$$\boxed{\Rightarrow \ell\ell(\theta) = -m \log Z(\theta) + \sum_{i=1}^{m} F_\theta(X^i)}$$

MRF log-likelihood - Exponential form

$$\Rightarrow \boxed{\ell\ell(\theta) = -m \log Z(\theta) + \sum_{i=1}^{m} F_\theta(X^i)}$$

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

$$\theta^* = \operatorname{argmax} \ell\ell(\theta)$$

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

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

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

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

Positive and Negative forces

K. N. Toosi
University of Technology

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

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

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

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

$$= m\left(\underbrace{E_P\left\{ \frac{\partial}{\partial \theta_k} F_\theta(X^i) \right\}}_{\text{positive phase}} - \underbrace{E_{P_\theta(X)}\left\{ \frac{\partial}{\partial \theta_k} F_\theta(X) \right\}}_{\text{negative phase}} \right)$$

Log-linear models

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

pgm 23

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

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

log-linear MRF / No shared parameters

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

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

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

log-linear models

log-linear $\quad \phi_{\theta_1}(A,B) = \exp\left(\sum_{k=1}^{p} w_k f_k(A,B)\right) \qquad \theta_1 = (w_1, \dots, w_p)$

$$\psi_{\theta_2}(B,C) = \exp\left(\sum_{k=1}^{q} u_k g_k(B,C)\right) \qquad \theta_2 = (u_1, \dots, u_q)$$

$$\phi_w(A,B) = \exp(w\, f(A,B))$$

$$\psi_u(B,C) = \exp(u\, g(B,C)) \Rightarrow \boxed{P(A,B,C) = \frac{1}{Z(w,u)} e^{w f(A,B) + u\, g(B,C)}}$$

$$ll(\theta) = -m \log\left(\sum_A\sum_B\sum_C e^{w f(A,B) + u g(B,C)}\right) + \sum_{i=1}^{m} w\, f(A^i,B^i) + \sum_{i=1}^{m} u\, g(B^i,C^i)$$

$$= -m \log\left(\sum_A\sum_B\sum_C e^{w f(A,B) + u g(B,C)}\right) + w \underbrace{\sum_{i=1}^{m} f(A^i,B^i)} + u \underbrace{\sum_{i=1}^{m} g(B^i,C^i)}$$

sufficient statistics

$$\frac{\partial ll(w,u)}{\partial w} = -m \frac{\sum_A\sum_B\sum_C f(A,B)\, e^{w f(A,B) + u g(B,C)}}{\boxed{\sum_A\sum_B\sum_C e^{w f(A,B) + u g(B,C)}}} + \sum_{i=1}^{m} f(A^i,B^i)$$

$$Z(\theta) = Z(w,u)$$

log-linear models

K. N. Toosi
University of Technology

29 II

$$\Rightarrow \frac{\partial \ell\ell(\theta)}{\partial w} = m\left(\frac{1}{m}\sum_{i=1}^{m} f(A^i,B^i) - \sum_A\sum_B\sum_C f(A,B)\underbrace{\frac{1}{Z(\theta)}\, e^{w f(A,B)+u g(B,C)}}_{\frac{1}{Z(\theta)}\Phi_w(A,B)\,\Psi_u(B,C)}\right)$$

$$= m\left(\frac{1}{m}\sum_{i=1}^{m} f(A^i,B^i) - \sum_A\sum_B\sum_C f(A,B)\, P_\theta(A,B,C)\right)$$

$$\frac{\partial \ell\ell(\theta)}{\partial w} = m\left(\mathcal{E}_D\{f(A^i,B^i)\} - \mathcal{E}_\theta\{f(A,B)\}\right)$$

$$\frac{\partial \ell\ell(\theta)}{\partial u} = m\left(\mathcal{E}_D\{g(B^i,C^i)\} - \mathcal{E}_\theta\{g(B,C)\}\right)$$

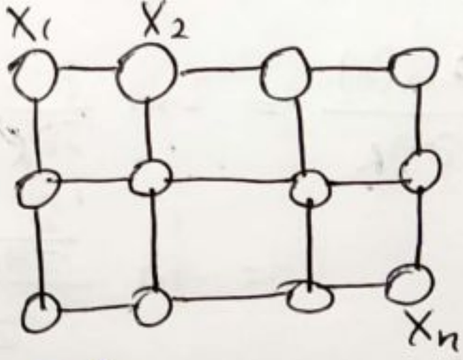
Example: Pairwise MRFs

K. N. Toosi
University of Technology

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:

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

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

$\vdots$

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

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

Example: Pairwise MRFs

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

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

Data:

$$X^1 = (X_1^1, X_2^1, \ldots, X_n^1)$$
$$X^2 = (X_1^2, X_2^2, \ldots, X_n^2)$$
$$\vdots$$
$$X^m = (X_1^m, X_2^m, \ldots, X_n^m)$$

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

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

$$= \;〃\; - m \sum_{X_1}\sum_{X_2}\cdots\sum_{X_m} P_\theta(X_1, X_2, \ldots X_m)\, f_i(X_i)$$

$$= \;〃\; - m \sum_{X_i}\Big[\sum_{X_1}\sum_{X_2}\cdots\sum_{X_{i-1}}\sum_{X_{i+1}}\cdots\sum_{X_n} P_\theta(X_1 \ldots X_m)\Big] f_i(X_i)$$

$$= \;〃\; - m \sum_{X_i} P_\theta(X_i)\, f(X_i) \longrightarrow \text{marginal distr}$$

Example: Pairwise MRFs

K. N. Toosi
University of Technology

pgm 23 (IV)

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

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

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

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

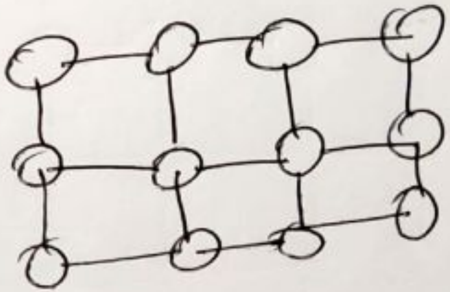
$P_\theta(X_i, X_j) \quad \forall (i,j) \in \mathcal{E}$

Example: Pairwise MRFs

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

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

Need inference
exact, approximate



Optimize Log-linear models

K. N. Toosi
University of Technology

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

$$\frac{\partial \ell\ell(\theta)}{\partial \theta_c} = \sum_{k=1}^{m} f(X_c^k) - m \sum_{X_c} \underbrace{P_\theta(X_c) f(X_c)}_{\text{need inferece}}$$

init $\theta = \theta_0$

while not converged

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

Learning MRFs, log-linear models

$$H_{ij} = \frac{\partial \ln(Z)}{\partial \theta_i \, \theta_j}$$

positive definite $u^T H u > 0$ for all $u \neq 0$

convex function

$\Rightarrow$ single unique global minimum

$-\ln(Z)$ concave

How to compute?

$$\mathcal{E}_\theta\{f(A,B)\} = \sum_A \sum_B \sum_C f(A,B)\, P_\theta(A,B,C)$$

$$= \sum_A \sum_B f(A,B) \sum_C P_\theta(A,B,C)$$

$$= \sum_A \sum_B f(A,B)\; P(A,B)$$

$\rightarrow$ Marginal distribut

large no. of variables

Learning MRFs, log-linear models

K. N. Toosi
University of Technology

$-\ln(Z)$ concave

How to compute?

$$\mathcal{E}_\theta\{f(A,B)\} = \sum_A \sum_B \sum_C f(A,B)\, P_\theta(A,B,C)$$

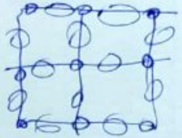
$$= \sum_A \sum_B f(A,B) \sum_C P_\theta(A,B,C)$$

$$= \sum_A \sum_B f(A,B) \quad P(A,B)$$

Marginal distribut

large no. of variables

$$\mathcal{E}_\theta\{f(A,B)\} = \sum_A \sum_B \cdots \sum_Z f(A,B)\, P_\theta(A,B,\dots,Z)$$



$$= \sum_A \sum_B f(A,B)\, P(A,B)$$

compute using inference (exact / Approximate)

VE, Junction tree

Loopy BP / Variational infere

Sample-Based (MCMC)

Learning MRFs - General Case

30

MRF

$$P_\theta(X) = P_\theta(X_1 \text{---} X_n) = \frac{1}{Z(\theta)} \prod_{c \in \mathcal{C}} \phi_c^\theta(X_c)$$

$$Z(\theta) = \sum_{X_1} \sum_{X_2} \cdots \sum_{X_n} \prod_{c \in \mathcal{C}} \phi_c^\theta(X_c)$$

$$ll(\theta) = -m \log Z(\theta) + \sum_{c \in \mathcal{C}} \sum_{i=1}^{m} \log \phi_c^\theta(X_c^i)$$

Challenges

* $Z(\theta)$ couples the parameters
* $Z(\theta)$ is a sum of an exponentially many terms. ~~large no. of terms~~

Recap

$$p_{\boldsymbol{\theta}}(\mathsf{X}) = \frac{1}{Z(\boldsymbol{\theta})} \exp(F(\mathsf{X}, \boldsymbol{\theta}))$$

$$\mathcal{D} = \{\mathsf{X}^1, \mathsf{X}^2, \ldots, \mathsf{X}^m\}$$

$$\ell\ell(\boldsymbol{\theta}) = \sum_{i=1}^{m} F(\mathsf{X}^i, \boldsymbol{\theta}) - m \ln(Z(\boldsymbol{\theta}))$$

Recap

$$p_{\boldsymbol{\theta}}(\mathsf{X}) = \frac{1}{Z(\boldsymbol{\theta})} \exp(F(\mathsf{X}, \boldsymbol{\theta})) \qquad \mathcal{D} = \{\mathsf{X}^1, \mathsf{X}^2, \ldots, \mathsf{X}^m\}$$

$$\ell\ell(\boldsymbol{\theta}) = \sum_{i=1}^{m} F(\mathsf{X}^i, \boldsymbol{\theta}) - m \ln(Z(\boldsymbol{\theta}))$$

$$\frac{\partial}{\partial \theta_j} \ell\ell(\boldsymbol{\theta}) = \sum_{i=1}^{m} \frac{\partial}{\partial \theta_j} F(\mathsf{X}^i, \boldsymbol{\theta}) - m \sum_{\mathsf{X}} p_{\boldsymbol{\theta}}(\mathsf{X}) \frac{\partial}{\partial \theta_j} F(\mathsf{X}, \boldsymbol{\theta})$$

$$= m\, \mathbb{E}_{\mathcal{D}} \left\{ \frac{\partial}{\partial \theta_j} F(\mathsf{X}, \boldsymbol{\theta}) \right\} - m\, \mathbb{E}_{p_{\boldsymbol{\theta}}(\mathsf{X})} \left\{ \frac{\partial}{\partial \theta_j} F(\mathsf{X}, \boldsymbol{\theta}) \right\}$$

Recap

$$F(\mathrm{X}, \boldsymbol{\theta}) = \sum_k \theta_k \; f_k(\mathrm{X}_{c_k}) \qquad\qquad \mathrm{X}_{c_k} \subset \mathrm{X}$$

$$\begin{aligned}
\frac{\partial}{\partial \theta_j} \ell\ell(\boldsymbol{\theta}) &= \sum_{i=1}^{m} \frac{\partial}{\partial \theta_j} F(\mathrm{X}^i, \boldsymbol{\theta}) - m \sum_{\mathrm{X}} p_{\boldsymbol{\theta}}(\mathrm{X}) \; \frac{\partial}{\partial \theta_j} F(\mathrm{X}, \boldsymbol{\theta}) \\
&= m \; \mathbb{E}_{\mathcal{D}} \left\{ \frac{\partial}{\partial \theta_j} F(\mathrm{X}, \boldsymbol{\theta}) \right\} - m \; \mathbb{E}_{p_{\boldsymbol{\theta}}(\mathrm{X})} \left\{ \frac{\partial}{\partial \theta_j} F(\mathrm{X}, \boldsymbol{\theta}) \right\} \\
&= m \; \mathbb{E}_{\mathcal{D}} \left\{ f_j(\mathrm{X}_{c_j}) \right\} - m \; \mathbb{E}_{p_{\boldsymbol{\theta}}(\mathrm{X})} \left\{ f_j(\mathrm{X}_{c_j}) \right\} \\
&= \sum_{i=1}^{m} f_j(\mathrm{X}^i_{c_j}) - m \sum_{\mathrm{X}} p_{\boldsymbol{\theta}}(\mathrm{X}) f_j(\mathrm{X}_{c_j}) \\
&= \sum_{i=1}^{m} f_j(\mathrm{X}^i_{c_j}) - m \sum_{\mathrm{X}_{c_j}} p_{\boldsymbol{\theta}}(\mathrm{X}_{c_j}) \; f_j(\mathrm{X}_{c_j})
\end{aligned}$$

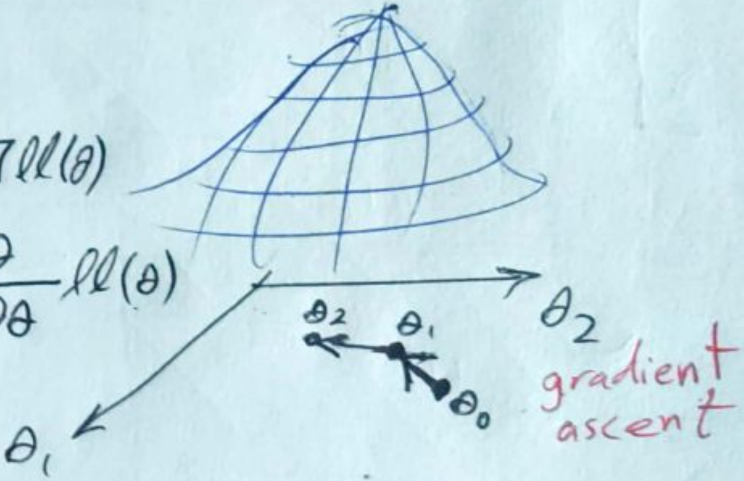
Gradient Ascent

$$\frac{\partial}{\partial \theta_j} ll(\theta)$$

$$P(X) = \frac{1}{Z(\theta)} e^{F(X,\theta)}$$

$$\begin{bmatrix} \frac{\partial}{\partial \theta_1} ll(\theta) \\ \frac{\partial}{\partial \theta_2} ll(\theta) \\ \vdots \\ \frac{\partial}{\partial \theta_p} ll(\theta) \end{bmatrix} \Rightarrow \nabla ll(\theta) = \frac{\partial}{\partial \theta} ll(\theta)$$

گرادیان



$$\frac{\partial}{\partial \theta_j} ll(\theta) = m\, E_D\left\{\frac{\partial}{\partial \theta_j} F(X,\theta)\right\} - m\, E_{P_\theta(X)}\left\{\frac{\partial}{\partial \theta_j} F(X,\theta)\right\}$$

Gradient Ascent

$$\frac{\partial}{\partial\theta_j} ll(\theta) = m\, E_D\left\{\frac{\partial}{\partial\theta_j} F(X,\theta)\right\} - m\, E_{P_\theta(X)}\left\{\frac{\partial}{\partial\theta_j} F(X,\theta)\right\}$$

$t \leftarrow 0$

θ_0

while not converged

$$\left.\frac{\partial}{\partial\theta} ll(\theta)\right|_{\theta=\theta_t}$$

$$\theta_{t+1} = \theta_t + \lambda \left.\frac{\partial}{\partial\theta} ll(\theta)\right|_{\theta=\theta_t}$$

$t \leftarrow t+1$

Learning By Sampling

K. N. Toosi
University of Technology

problem: How to compute $E_{P_\theta(X)}\left\{\frac{\partial}{\partial\theta_j} F(X,\theta)\right\}$

take samples $X'^1, X'^2, \dots, X'^M$ from $P_\theta(X) = \frac{1}{Z(\theta)} e^{F(X,\theta)}$

$$E_{P_\theta(X)}\left\{\frac{\partial}{\partial\theta_j} F(X,\theta)\right\} \simeq \frac{1}{M} \sum_{i=1}^{m} \frac{\partial}{\partial\theta_j} F(X'^i,\theta)$$

$$\frac{\partial}{\partial\theta_j} ll(\theta) = \sum_{i=1}^{m} \frac{\partial}{\partial\theta_j} F(X^i,\theta) - \frac{m}{M} \sum_{i=1}^{m} \frac{\partial}{\partial\theta_j} F(X'^i,\theta)$$

$\boxed{\text{usually } M=m}$

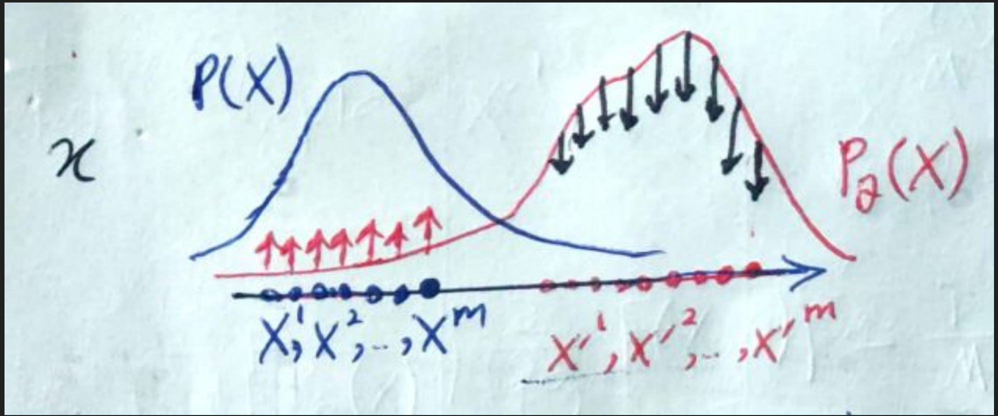
training data $X^1, X^2, \dots, X^m$

samples from $P_\theta(X)$ $X'^1, X'^2, \dots, X'^m$

$$\frac{\partial}{\partial\theta_j} ll(\theta) = \underbrace{\sum_{i=1}^{m} \frac{\partial F}{\partial\theta_j} F(X^i,\theta)}_{\text{positive force}} - \underbrace{\sum_{i=1}^{m} \frac{\partial F}{\partial\theta_j} F(X'^i,\theta)}_{\text{negative force}}$$

Positive and Negative Forces

$$\frac{\partial}{\partial \theta_j} \ell\ell(\theta) = \underbrace{\sum_{i=1}^{m} \frac{\partial F}{\partial \theta_j} F(X^i, \theta)}_{\text{positive force}} - \underbrace{\sum_{i=1}^{m} \frac{\partial F}{\partial \theta_j} F(X'^i, \theta)}_{\text{negative force}}$$



MCMC-based learning

while not converged **do**
  Sample a minibatch of m examples $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\}$ from the training set
  $\mathbf{g} \leftarrow \frac{1}{m} \sum_{i=1}^{m} \nabla_{\boldsymbol{\theta}} \log \tilde{p}(\mathbf{x}^{(i)}; \boldsymbol{\theta})$.
  Initialize a set of m samples $\{\tilde{\mathbf{x}}^{(1)}, \ldots, \tilde{\mathbf{x}}^{(m)}\}$ to random values (e.g., from a uniform or normal distribution, or possibly a distribution with marginals matched to the model's marginals).
  for $i = 1$ to k **do**
    for $j = 1$ to m **do**
      $\tilde{\mathbf{x}}^{(j)} \leftarrow \text{gibbs\_update}(\tilde{\mathbf{x}}^{(j)})$.
    end for
  end for
  $\mathbf{g} \leftarrow \mathbf{g} - \frac{1}{m} \sum_{i=1}^{m} \nabla_{\boldsymbol{\theta}} \log \tilde{p}(\tilde{\mathbf{x}}^{(i)}; \boldsymbol{\theta})$.
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \epsilon \mathbf{g}$.
end while

Goodfellow et al. "Deep Learning," MIT Press, 2016, Chapter 18

Contrastive Divergence

while not converged **do**
  Sample a minibatch of m examples $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\}$ from the training set
  $\mathbf{g} \leftarrow \frac{1}{m} \sum_{i=1}^{m} \nabla_{\boldsymbol{\theta}} \log \tilde{p}(\mathbf{x}^{(i)}; \boldsymbol{\theta})$.
  for $i = 1$ to m **do**
    $\tilde{\mathbf{x}}^{(i)} \leftarrow \mathbf{x}^{(i)}$.
  end for
  for $i = 1$ to k **do**
    for $j = 1$ to m **do**
      $\tilde{\mathbf{x}}^{(j)} \leftarrow \text{gibbs\_update}(\tilde{\mathbf{x}}^{(j)})$.
    end for
  end for
  $\mathbf{g} \leftarrow \mathbf{g} - \frac{1}{m} \sum_{i=1}^{m} \nabla_{\boldsymbol{\theta}} \log \tilde{p}(\tilde{\mathbf{x}}^{(i)}; \boldsymbol{\theta})$.
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \epsilon \mathbf{g}$.
end while

Goodfellow et al. "Deep Learning," MIT Press, 2016, Chapter 18

K. N. Toosi
University of Technology

Persistent Contrastive Divergence

Initialize a set of m samples $\{\tilde{\mathbf{x}}^{(1)}, \ldots, \tilde{\mathbf{x}}^{(m)}\}$ to random values (e.g., from a uniform or normal distribution, or possibly a distribution with marginals matched to the model's marginals).

while not converged **do**

  Sample a minibatch of m examples $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\}$ from the training set

  $\mathbf{g} \leftarrow \frac{1}{m} \sum_{i=1}^{m} \nabla_{\boldsymbol{\theta}} \log \tilde{p}(\mathbf{x}^{(i)}; \boldsymbol{\theta})$.

  for $i = 1$ to k **do**

    for $j = 1$ to m **do**

      $\tilde{\mathbf{x}}^{(j)} \leftarrow \text{gibbs\_update}(\tilde{\mathbf{x}}^{(j)})$.

    end for

  end for

  $\mathbf{g} \leftarrow \mathbf{g} - \frac{1}{m} \sum_{i=1}^{m} \nabla_{\boldsymbol{\theta}} \log \tilde{p}(\tilde{\mathbf{x}}^{(i)}; \boldsymbol{\theta})$.

  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \epsilon \mathbf{g}$.

end while

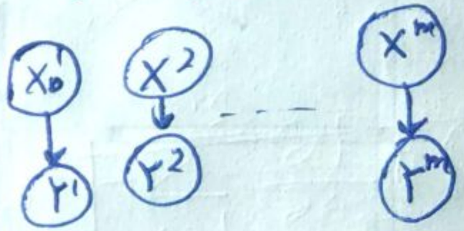
Goodfellow et al. "Deep Learning," MIT Press, 2016, Chapter 18

Learning CRFs - Conditional Likelihood

CRF - $P_\theta(Y|X) = \frac{1}{Z(\theta,X)} e^{F_\theta(X,Y)}$

Data

$(X^1,Y^1),(X^2,Y^2),\dots,(X^m,Y^m)$



Conditional likelihood

$cll(\theta) = Pr(Y^1,Y^2,\dots,Y^m|X^1,X^2,\dots,X^m)$

$\prod_{i=1}^{m} Pr(Y^i|X^1,\dots,X^m)$

$\prod_{i=1}^{m} Pr(Y^i|X^i) \Longrightarrow cll(\theta)=\prod_{i=1}^{m} P_\theta(Y^i|X^i)$

Learning CRFs

K. N. Toosi
University of Technology

TV

$$cll(\theta) = \log cl(\theta) = \sum_{i=1}^{m} \log P_\theta(Y^i \mid X^i)$$

$$= \sum_{i=1}^{m} \log \frac{1}{Z_\theta(X^i)} e^{F_\theta(X^i; Y^i)}$$

$$= -\sum_{i=1}^{m} \log Z_\theta(X^i) + \sum_{i=1}^{m} F_\theta(X^i; Y^i)$$

$$\frac{\partial}{\partial \theta_j} cll(\theta) = -\sum_{i=1}^{m} \sum_{Y} P_\theta(Y \mid X^i)\, F_\theta(X^i; Y) + \sum_{i=1}^{m} F_\theta(X^i; Y^i)$$

$$-\sum_{i=1}^{m} \underbrace{E_{P_\theta(Y \mid X^i)} \left\{ F_\theta(X^i; Y) \right\}}_{\text{Needs inference per } X^i} + m\, E_D \left\{ F_\theta(X, Y) \right\}$$

Learning CRFs

Conditional log-likelihood

$$\ell\ell(\theta) = \sum_{i=1}^{m} \log P_\theta(Y^i | X^i) = \sum_{i=1}^{m} \left(-\log Z_\theta(X^i) + \sum_{c \in C} \log \phi_c^\theta(Y_c^i, X^i) \right)$$

$$= -\sum_{i=1}^{m} \log Z_\theta(X^i) + \sum_{c \in C} \sum_{i=1}^{m} \log \phi_c^\theta(Y_c^i, X^i)$$

Log-linear models $\theta = (w_1, w_2, \ldots)$

$Y_{j} = \{Y_i\}_{i \in j}$

$$P_\theta(Y | X) = \frac{1}{Z(\theta, X)} \exp\left(\sum_{j \in J} w_j f_j(X, Y_j) \right)$$

$$\ell\ell(\theta) = -\sum_{i=1}^{m} \log Z(\theta, X^i) + \sum_{j \in J} w_j \sum_{i=1}^{m} f_j(X^i, Y^i)$$

Learning CRFs

$$ll(\theta) = -\sum_{i=1}^{m} \log Z(\theta, X^i) + \sum_{j \in J} w_j \sum_{i=1}^{m} f_j(X^i; Y^i)$$

$$\frac{\partial}{w_k} ll(\theta) = -\sum_{i=1}^{m} \sum_{Y} f_k(X^i; Y)\, P(Y \mid X^i) + \sum_{i=1}^{m} f_k(X^i, Y^i)$$

$$= -\sum_{i=1}^{m} \underbrace{E}_{P_\theta(Y \mid X^i)} \left\{ f_k(X^i; Y) \right\} + m\, E_D \left\{ f_k(X^i; Y^i) \right\}$$

Example

$$\sum_{Y} f_k(X^i, Y_3, Y_5)\, P(Y \mid X^i) = \sum_{Y_3} \sum_{Y_5} f_k(X^i; Y_3, Y_5)\, \underbrace{P(Y_3, Y_5 \mid X^i)}_{\text{inference}}$$

Learning with Shared Parameters

K. N. Toosi
University of Technology

$$F_\theta(X) = \theta\left\{f_1(X_1, X_2) + f_2(X_2, X_3) + f_3(X_3, X_4)\right\}$$

$$E\left\{\frac{\partial F}{\partial \theta}\right\} = E\left\{f_1(X_1, X_2) + f_2(X_2, X_3) + f_3(X_3, X_4)\right\}$$

$$E\left\{f_1(X_1, X_2)\right\} + E\left\{f_2(X_2, X_3)\right\} + E\left\{f_3(X_3, X_4)\right.$$

References

- Goodfellow et al. "Deep Learning," MIT Press, 2016, Chapter 18
- Nowozin, et al. "Structured learning and prediction in computer vision," 2011, Chapter 5
- Koller, Daphne, and Nir Friedman. Probabilistic graphical models: principles and techniques. MIT press, 2009

K. N. Toosi University of Technology