Parameter Learning in MN

Amr Ahmed

10708-F08 Recitation

Outline

- CRF
- Learning CRF for 2-d image segmentation
- IPF parameter sharing revisited

Log-linear Markov network (most common representation)

- **Feature** is some function $\phi[\mathbf{D}]$ for some subset of variables **D**
  - e.g., indicator function
- **Log-linear model** over a Markov network *H*:
  - a set of features $\phi_1[\mathbf{D}_1],\ldots, \phi_k[\mathbf{D}_k]$
    - each $\mathbf{D}_i$ is a subset of a clique in *H*
    - two ϕ's can be over the same variables
  - a set of weights $w_1,\ldots,w_k$
    - usually learned from data
  - $$P(X) = \frac{1}{Z}\exp\left[\sum_i w_i \phi_i(D_i)\right]$$

Generative v. Discriminative classifiers – A review

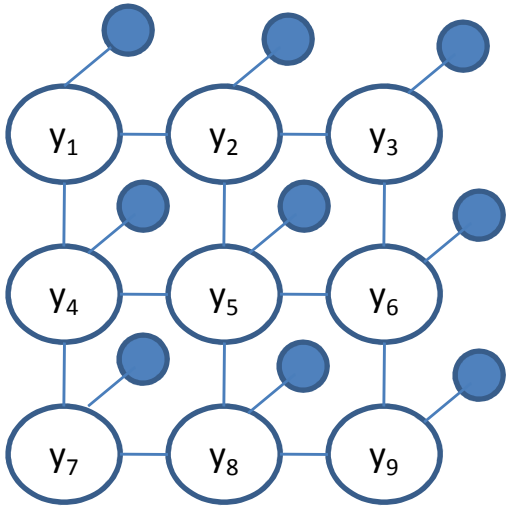
- **Want to Learn**: h:**X** $\mapsto$ Y
  - **X** – features
  - Y – target classes
- **Bayes optimal classifier** – P(Y|**X**)
- **Generative classifier**, e.g., Naïve Bayes:
  - Assume some **functional form for P(X|Y), P(Y)**
  - Estimate parameters of P(**X**|Y), P(Y) directly from training data
  - Use Bayes rule to calculate P(Y|**X= x**)
  - This is a **'*generative*' model**
    - **Indirect** computation of P(Y|**X**) through Bayes rule
    - But, **can generate a sample of the data**, P(**X**) = $\sum_y$ P(y) P(**X**|y)
- **Discriminative classifiers**, e.g., Logistic Regression:
  - Assume some **functional form for P(**Y**|X)**
  - Estimate parameters of P(Y|**X**) directly from training data
  - This is the **'*discriminative*' model**
    - Directly learn P(Y|**X**)
    - But **cannot obtain a sample of the data**, because P(**X**) is not available

Log-linear CRFs (most common representation)

- **Graph *H*:** only over hidden vars $Y_1,..,Y_P$
  - **No assumptions about dependency on observed vars *X***
  - You must always observe all of **X**
- **Feature** is some function $\phi[\mathbf{D}]$ for some subset of variables **D**
  - e.g., indicator function,
- **Log-linear model** over a CRF *H*:
  - a set of features $\phi_1[\mathbf{D}_1],\ldots, \phi_k[\mathbf{D}_k]$
    - each $\mathbf{D}_i$ is a subset of a clique in *H*
    - two ϕ's can be over the same variables
  - a set of weights $w_1,\ldots,w_k$
    - usually learned from data
  - $$P(Y \mid X) = \frac{1}{Z(X)} \exp\left[\sum_i w_i \phi_i(D_i, X)\right]$$

Example: Image Segmentation

- A set of features $\phi_1[\mathbf{D}_1],\ldots, \phi_k[\mathbf{D}_k]$
  - each $\mathbf{D}_i$ is a subset of a clique in *H*
  - two ϕ's can be over the same variables



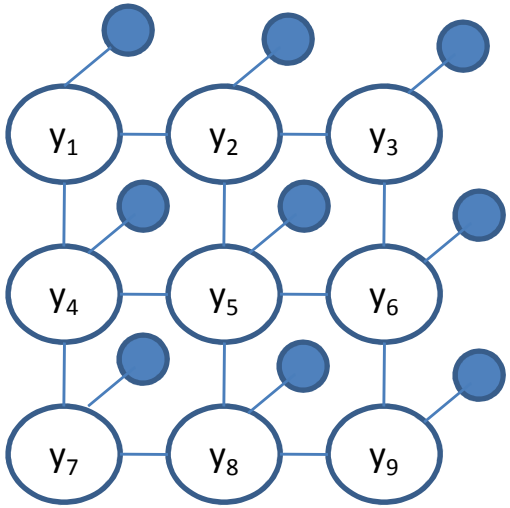
$$P(Y \mid X) = \frac{1}{Z(X)} \exp\left[\sum_i w_i \phi_i(D_i, X)\right]$$

We will define features as follows:

- $\phi(x, y)$: measures compatibility of node color and its segmentation
- A set of indicator features triggered for each edge labeling pair {ff,bb,fb,bf}
  - This is a allowed since we can define many features overr the same subset of variables

Example: Image Segmentation

- A set of features $\phi_1[\mathbf{D}_1],\ldots, \phi_k[\mathbf{D}_k]$
  - each $\mathbf{D}_i$ is a subset of a clique in *H*
  - two ϕ's can be over the same variables



$$P(Y \mid X) = \frac{1}{Z(X)} \exp\left[\sum_i w_i \phi_i(D_i, X)\right]$$

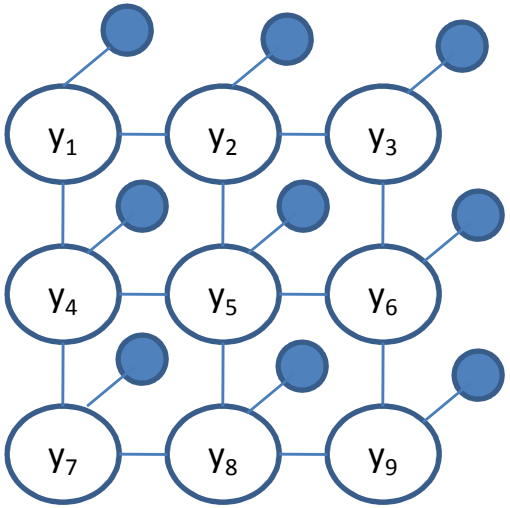
$$\phi(x, y) = \begin{cases} \log P\left(x \mid GMM_b\right) & y = b \\ \log P\left(x \mid GMM_f\right) & y = f \end{cases}$$

$$\phi_{ff}(y_i, y_j) = \begin{cases} 1 & y_i = f, y_j = f \\ 0 & otherwise \end{cases} \qquad \phi_{fb}(y_i, y_j) = \begin{cases} 1 & y_i = f, y_j = b \\ 0 & otherwise \end{cases}$$

$$\phi_{bf}(y_i, y_j) = \begin{cases} 1 & y_i = b, y_j = f \\ 0 & otherwise \end{cases} \qquad \phi_{bb}(y_i, y_j) = \begin{cases} 1 & y_i = b, y_j = b \\ 0 & otherwise \end{cases}$$

Example: Image Segmentation

- A set of features $\phi_1[\mathbf{D}_1],\ldots, \phi_k[\mathbf{D}_k]$
  - each $\mathbf{D}_i$ is a subset of a clique in *H*
  - two ϕ's can be over the same variables



$$P(Y \mid X) = \frac{1}{Z(X)} \exp\left[\sum_i w_i \phi_i(D_i, X)\right]$$

-Now we just need to sum these features

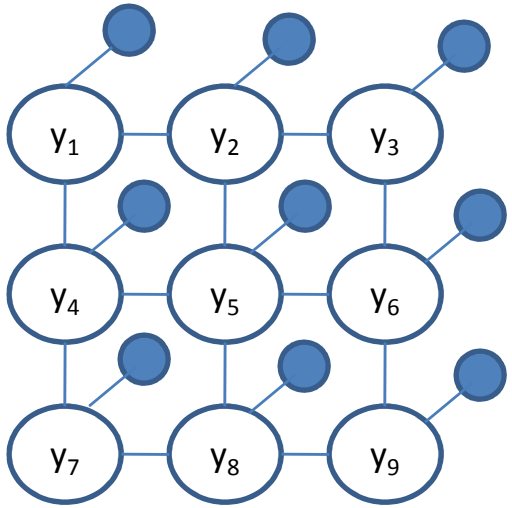
$$P(Y \mid X) \propto \exp\left[\sum_{i \in V} \phi(x_i, y_i) + \sum_{ij \in E} \sum_{m \in \{ff, fb, bf, bb\}} w_m \phi_m(y_i, y_j)\right]$$

$$P(Y \mid X) \propto \exp\left[\sum_{i \in V} \phi(x_i, y_i) + \sum_{m \in \{ff, fb, bf, bb\}} w_m C_m\right] \qquad C_m = \sum_{ij \in E} \phi_m(y_i, y_j) = \sum_{ij \in E} I(y_i y_j = m)$$

We need to learn parameters w_m

Example: Image Segmentation

$$P(Y \mid x) \propto \exp\left[\sum_{i \in V} \phi(x_i, y_i) + \sum_{m \in \{ff, fb, bf, bb\}} w_m C_m\right]$$



Given N data points (images and their segmentations)

$$\frac{\partial \ell(Data : w)}{\partial w_m} = \sum_{n=1}^{N} C_m[n] - E_w[C_m \mid X[n]]$$

$$C_m = \sum_{ij \in E} \phi_m(y_i, y_j) = \sum_{ij \in E} I(y_i y_j = m)$$

Count for features *m* in data n

Requires inference using the current parameter estimates

Example: Inference for Learning

$$P(Y \mid X) \propto \exp\left[\sum_{i \in V} f(x_i, y_i) + \sum_{m \in \{ff, fb, bf, bb\}} w_m C_m\right]$$

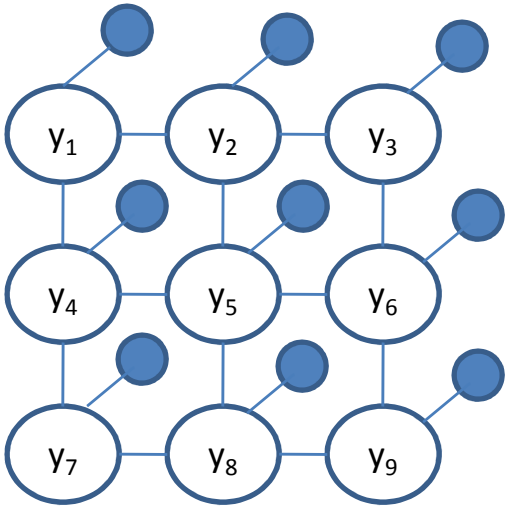
$$\frac{\partial \ell(Data : w)}{\partial w_m} = \sum_{n=1}^{N} C_m[n] - E_w\left[C_m \mid X[n]\right]$$

How to compute $E[C_{fb} \mid X[n]]$

$$E_w\left[C_{fb} \mid X[n]\right] = E\left[\sum_{ij} I(y_i = f, y_j = b) \mid X[n]\right]$$

$$E_w\left[C_{fb} \mid X[n]\right] = \sum_{ij} E_w\left[I(y_i = f, y_j = b) \mid X[n]\right]$$

$$E_w\left[C_{fb} \mid X[n]\right] = \sum_{ij} P_w\left(y_i = f, y_j = b \mid X[n]\right)$$



$$C_m = \sum_{ij \in E} \phi_m\left(y_i, y_j\right) = \sum_{ij \in E} I(y_i y_j = m)$$

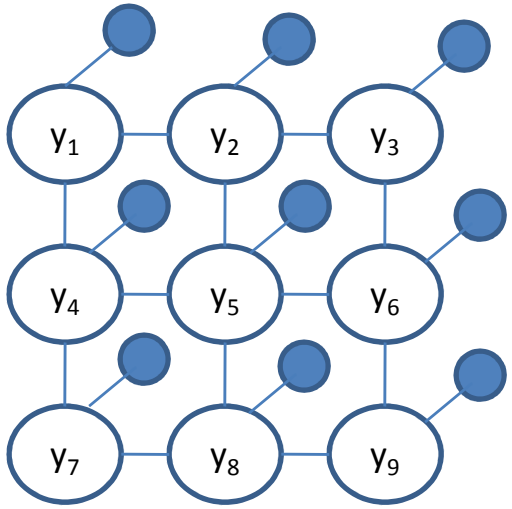
Example: Inference for Learning

$$P(Y \mid X) \propto \exp\left[\sum_{i \in V} f(x_i, y_i) + \sum_{m \in \{ff, fb, bf, bb\}} w_m C_m\right]$$

$$\frac{\partial \ell(Data : w)}{\partial w_m} = \sum_{n=1}^{N} C_m[n] - E_w[C_m \mid X[n]]$$

How to compute $E[C_{fb} \mid X[n]]$

$$E_w[C_{fb} \mid X[n]] = \sum_{ij} P_w(y_i = f, y_j = b \mid X[n])$$



$$C_m = \sum_{ij \in E} \phi_m(y_i, y_j) = \sum_{ij \in E} I(y_i y_j = m)$$

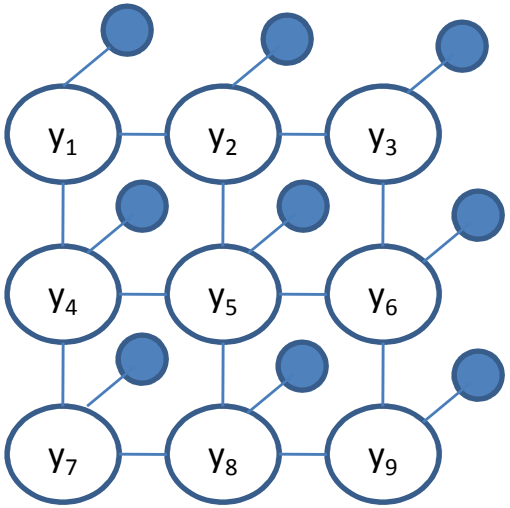
Representation Equivalence

Log linear representation

$$P(Y \mid X) \propto \exp\left[\sum_{i \in V} \phi(x_i, y_i) + \sum_{m \in \{ff, fb, bf, bb\}} w_m C_m\right]$$

$$C_m = \sum_{ij \in E} \phi_m(y_i, y_j) = \sum_{ij \in E} I(y_i y_j = m)$$

$$\phi(x, y) = \begin{cases} \log P(x \mid GMM_b) & y = b \\ \log P(x \mid GMM_f) & y = f \end{cases}$$



Tabular MN representation from HW4

$$P(Y \mid X) \propto \prod_{i \in V} \Phi(x_i, y_i) \prod_{ij \in E} \psi(y_i, y_j)$$

$$\prod_{i \in V} \Phi(x_i, y_i) = \prod_{i \in V} P(x_i \mid GMM_{y_i})$$

$$= \prod_{i \in V} \exp\left[\log P(x_i \mid GMM_{y_i})\right]$$

$$= \exp\left[\sum_{i \in V} \log P(x_i \mid GMM_{y_i})\right]$$

$$= \exp\left[\sum_{i \in V} \phi(x_i, y_i)\right]$$

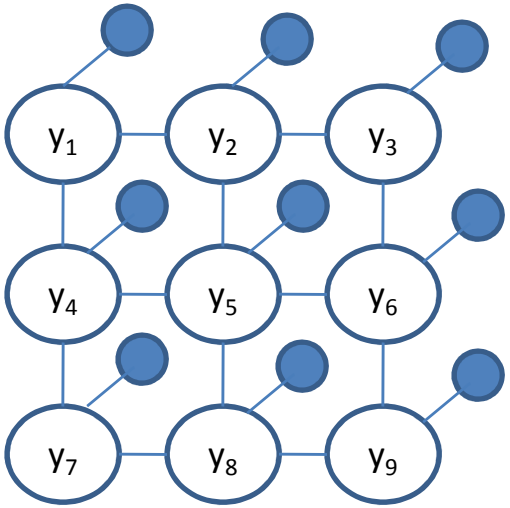
Representation Equivalence

Log linear representation

$$P(Y \mid X) \propto \exp\left[\sum_{i \in V} \phi(x_i, y_i) + \sum_{m \in \{ff, fb, bf, bb\}} w_m C_m\right]$$

$$C_m = \sum_{ij \in E} \phi_m(y_i, y_j) = \sum_{ij \in E} I(y_i y_j = m)$$

$$\phi(x, y) = \begin{cases} \log P(x \mid GMM_b) & y = b \\ \log P(x \mid GMM_f) & y = f \end{cases}$$



Tabular MN representation from HW4

Now do it over the edge potential

$$P(Y \mid X) \propto \prod_{i \in V} \Phi(x_i, y_i) \prod_{ij \in E} \psi(y_i, y_j)$$

$$\psi(y_i, y_j) = \theta_{ff}^{I(y_i y_j = ff)} \theta_{fb}^{I(y_i y_j = fb)} \theta_{bf}^{I(y_i y_j = bf)} \theta_{bb}^{I(y_i y_j = bb)}$$

$$\psi(y_i, y_j) = \prod_{m = \{ff, bf, fb, bb\}} \theta_m^{I(y_i y_j = m)}$$

This is correct as for every assignment to $y_i y_j$ we select one value from the table

Tabular MN representation from HW4

Now do it over the edge potential

$$P(Y \mid X) \propto \prod_{i \in V} \Phi(x_i, y_i) \prod_{ij \in E} \psi(y_i, y_j)$$

$$\psi(y_i, y_j) = \prod_{m=\{ff,bf,fb,bb\}} \theta_m^{I(y_i y_j = m)}$$

This is correct as for every assignment to $y_i y_j$ we select one value from the table

$$\psi(y_i, y_j) = \prod_{m=\{ff,bf,fb,bb\}} \theta_m^{I(y_i y_j = m)} = \prod_{m=\{ff,bf,fb,bb\}} \exp[\log \theta_m^{I(y_i y_j = m)}]$$

The cheap exp(log..) trick

$$\psi(y_i, y_j) = \exp\left[\sum_{m=\{ff,bf,fb,bb\}} I(y_i y_j = m) \log \theta_m\right]$$

Just algebra

Now lets combine it over all edge assuming parameter sharing

$$\prod_{ij \in E} \psi(y_i, y_j) = \prod_{ij \in E} \exp\left[\sum_{m=\{ff,bf,fb,bb\}} I(y_i y_j = m) \log \theta_m\right] = \exp\left[\sum_{ij \in E} \sum_{m=\{ff,bf,fb,bb\}} I(y_i y_j = m) \log \theta_m\right]$$

Now use the same C_m trick

$$\exp\left[\sum_{m=\{ff,bf,fb,bb\}} C_m \log \theta_m\right] \longleftarrow C_m = \sum_{ij \in E} I(y_i y_j = m)$$

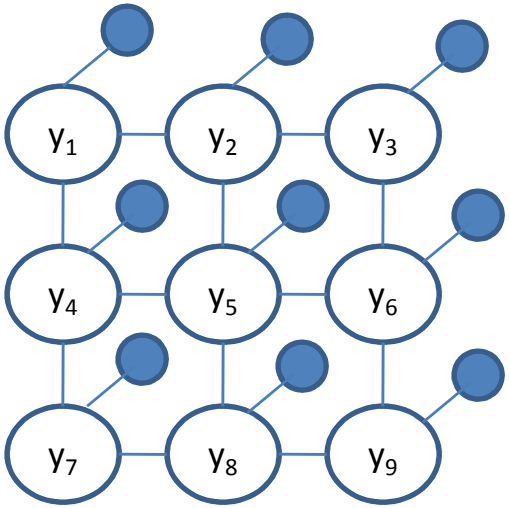
Representation Equivalence

Log linear representation

$$P(Y \mid X) \propto \exp\left[\sum_{i \in V} \phi(x_i, y_i) + \sum_{m \in \{ff, fb, bf, bb\}} w_m C_m\right]$$

$$C_m = \sum_{ij \in E} \phi_m(y_i, y_j) = \sum_{ij \in E} I(y_i y_j = m)$$

$$\phi(x, y) = \begin{cases} \log P(x \mid GMM_b) & y = b \\ \log P(x \mid GMM_f) & y = f \end{cases}$$



Tabular MN representation from HW4

$$P(Y \mid X) \propto \prod_{i \in V} \Phi(x_i, y_i) \prod_{ij \in E} \psi(y_i, y_j)$$

Now substitute

$$P(Y \mid X) \propto \exp\left[\sum_{i \in V} \phi(x_i, y_i) + \sum_{m \in \{ff, fb, bf, bb\}} C_m \log \theta_m\right]$$

Equivalent, with $w_m = \log \theta_m$ Where θ is the value in the tabular

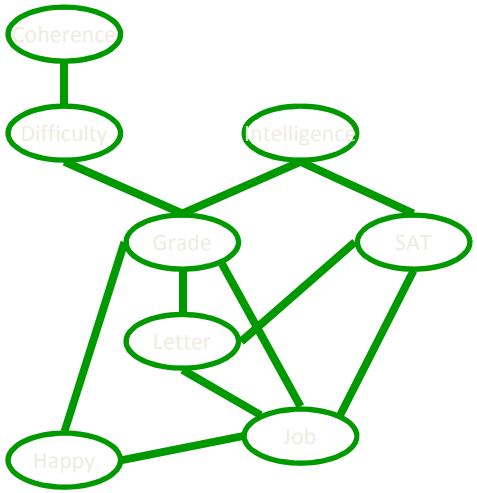
Outline

- CRF
- Learning CRF for 2-d image segmentation
- IPF parameter sharing revisited

Iterative Proportional Fitting (IPF)

$$\hat{P}(\mathbf{u}) = \frac{\text{Count}(\mathbf{U} = \mathbf{u})}{m}$$

$$\frac{\partial \ell}{\partial \psi_i(\mathbf{c}_i)} = \frac{m\hat{P}(\mathbf{c}_i)}{\psi_i(\mathbf{c}_i)} - \frac{mP^{\psi}_{\mathcal{F}}(\mathbf{c}_i)}{\psi_i(\mathbf{c}_i)} = 0$$



- Setting derivative to zero:

$$\frac{m\hat{P}(c_i)}{\psi_i(c_i)} - \frac{mP^{\psi}_{\mathcal{F}}(c_i)}{\psi_i(c_i)} = 0$$

- Fixed point equation:

$$\psi_i(c_i) = \psi_i(c_i)\frac{\hat{P}(c_i)}{P^{\psi}_{\mathcal{F}}(c_i)}$$

must initialize $\psi_i(c_i) > 0$

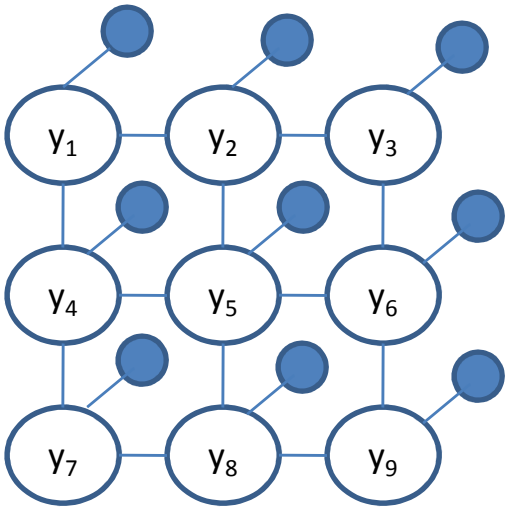
- Iterate and converge to optimal parameters
  - Each iteration, must compute:

$$\psi_i^{(t+1)}(c_i) \leftarrow \frac{\hat{P}(c_i)}{P^{\psi^{(t)}}_{\mathcal{F}}(c_i)}\psi_i^{(t)}(c_i)$$

compute $P^{\psi^{(t)}}_{\mathcal{F}}(c_i)$ at iteration t

Parameter Sharing in your HW

- Note that I am suing Y for label
- All edge potentials are shared
- Also we are learning a conditional model



IPF parameter sharing

We only have one data point (image) in this example so we dropped X[n] to only X

$$\psi_i^{t+1}(c_i) = \psi_i^t(c_i) \frac{\hat{p}(c_i \mid X)}{p(c_i \mid \psi^t, X)}$$

In total we have 4 parameters as opposed to 4 parameters per edge

How to calculate these quantities using parameter sharing?

$$\hat{p}(c = fb \mid X) = \frac{\sum_{ij \in E} I(y_i = f, y_j = b \mid X)}{|E|}$$

We can cancel |E| due to division

$$P(c = fb \mid \psi^t, X) = \frac{\sum_{ij \in E} P(y_i = f, y_j = b \mid \psi^t, X)}{|E|}$$

Run lbp ,when converged

$$P(y_i, y_j \mid \psi^t, X) \propto \Phi(x_i, y_i)\Phi(x_j, y_j)\psi(y_i, y_j) \prod_{k \in N(i)-j} \delta_{k \to i} \prod_{k' \in N(j)-i} \delta_{k' \to j}$$