Week 3: Tutorial

The intuition for how the Hammersley-Clifford theorem works

The goal of this short section is to give an intuition behind the Hammersley-Clifford theorem by explicitly showing that it holds for a particular example. Consider a simple chain $X - Y - Z$. The corresponding graphical model is given by all distributions that factorize

$$(*) \qquad f(x,y,z) = \alpha(x,y)\beta(y,z).$$

We want to show that this is equivalent to $X \perp Z | Y$ as long as $\alpha(x,y) > 0$ and $\beta(y,z) > 0$ for all x, y, z.

We will use the characterization that $X \perp Z | Y$ if and only if $f(x|y,z) = f(x|y)$ does not depend on z.

We first show that the conditional independence $X \perp Z | Y$ implies the particular factorization in $(*)$. Note that

$$f(x,y,z) = f(y,z)f(x|y,z) = f(x|y)f(y,z).$$

So the factorization in $(*)$ works with $\alpha(x,y) = f(x|y)$ and $\beta(y,z) = f(y,z)$.

Now we will show that the factorization in $(*)$ implies conditional independence. Indeed, note that $(*)$ implies that

$$f(y,z) = (\sum_x \alpha(x,y))\beta(y,z).$$

and so

$$f(x|y,z) = \frac{\alpha(x,y)\beta(y,z)}{(\sum_x \alpha(x,y))\beta(y,z)} = \frac{\alpha(x,y)}{\sum_x \alpha(x,y)}$$

which does not depend on z proving the conditional independence.

Gaussian log-likelihood

Suppose we observe some i.i.d. data $\mathbf{x}_{1:n} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ from the m-variate Gaussian distribution $N_m(\mu, \Sigma)$. The density is

$$f(\mathbf{x}; \mu, \Sigma) \;=\; \frac{1}{(2\pi)^{m/2}} (\det \Sigma)^{-1/2} \exp\{-\tfrac{1}{2}(\mathbf{x} - \mu)^\top \Sigma^{-1} (\mathbf{x} - \mu)\}.$$

It is convenient to equivalently express this density in terms of $K = \Sigma^{-1}$:

$$f(\mathbf{x};\mu,K) \;=\; \frac{1}{(2\pi)^{m/2}}(\det(K))^{1/2}\exp\{-\tfrac{1}{2}(\mathbf{x}-\mu)^\top K(\mathbf{x}-\mu)\},$$

after taking logarithms it becomes

$$\log f(\mathbf{x};\mu,K) = -\frac{m}{2}\log(2\pi) + \frac{1}{2}\log\det K - \frac{1}{2}(\mathbf{x}-\mu)^\top K(\mathbf{x}-\mu).$$

Up to the obvious constants that do not depend on μ and K, the log-likelihood is

$$\ell_n(\mu,K) \;=\; \sum_{i=1}^n \log f(\mathbf{x}_i;\mu,K) \;=\; (\text{const}) + \frac{n}{2}\log\det(K) - \frac{1}{2}\sum_{i=1}^n(\mathbf{x}_i-\mu)^\top K(\mathbf{x}_i-\mu).$$

Irrespective of the value of K, the optimal $\hat{\mu}$ satisfies

$$\hat{\mu} \;=\; \bar{\mathbf{x}}_n = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i$$

This is because the gradient of $\nabla_\mu \ell_n$ is

$$\nabla_\mu \ell_n(\mu,K) \;=\; -\frac{1}{2}\sum_{i=1}^n(2K\mu - 2K\mathbf{x}_i) \;=\; -nK\mu + K\sum_{i=1}^n \mathbf{x}_i \;=\; nK(\bar{\mathbf{x}}_n - \mu).$$

Since K is invertible, this can be zero if and only if $\mu = \bar{\mathbf{x}}_n$.
We can thus consider the profile likelihood

$$\ell_n(\bar{\mathbf{x}}_n,K) \;=\; (\text{const}) + \frac{n}{2}\log\det(K) - \frac{1}{2}\sum_{i=1}^n(\mathbf{x}_i-\bar{\mathbf{x}}_n)^\top K(\mathbf{x}_i-\bar{\mathbf{x}}_n).$$

Note that

$$\begin{aligned}
\sum_{i=1}^n(\mathbf{x}_i-\bar{\mathbf{x}}_n)^\top K(\mathbf{x}_i-\bar{\mathbf{x}}_n) &= \sum_{i=1}^n \operatorname{tr}((\mathbf{x}_i-\bar{\mathbf{x}}_n)^\top K(\mathbf{x}_i-\bar{\mathbf{x}}_n)) \\
&= \sum_{i=1}^n \operatorname{tr}(K(\mathbf{x}_i-\bar{\mathbf{x}}_n)(\mathbf{x}_i-\bar{\mathbf{x}}_n)^\top) \\
&= n\operatorname{tr}\left(K\left\{\frac{1}{n}\sum_{i=1}^n(\mathbf{x}_i-\bar{\mathbf{x}}_n)(\mathbf{x}_i-\bar{\mathbf{x}}_n)^\top\right\}\right) \\
&= n\operatorname{tr}(KS_n),
\end{aligned}$$

where S_n is the sample covariance matrix. Note that $\bar{\mathbf{x}}_n$ and S_n form the sufficient statistics for the Gaussian model. With this new notation

$$\ell_n(\bar{\mathbf{x}}_n,K) = (\text{const}) + \frac{n}{2}(\log\det(K) - \operatorname{tr}(KS_n)).$$

Some useful facts:

- $\log\det(K)$ is a strictly concave function of K.
- $\mathrm{tr}(KS_n)$ is linear in K.
- The gradients are $\nabla_K \log\det(K) = K^{-1} = \Sigma$ and $\nabla_K \mathrm{tr}(KS_n) = S_n$.
- The MLE is $\hat{\Sigma} = S_n$ (this is where the gradient vanishes).

MRFs as exponential families

Consider a simple undirected graph $X_1 - X_2 - X_3$ where each variable is binary. Consider the following graphical model

$$p(x_1, x_2, x_3|\theta) \;=\; \frac{1}{Z(\theta)}\psi_{1,2}(x_1, x_2|\theta_{1,2})\psi_{2,3}(x_2, x_3|\theta_{2,3})$$

or equivalently

$$p(x_1, x_2, x_3|\theta) \;=\; \exp\Big\{\log\psi_{1,2}(x_1, x_2|\theta_{1,2}) + \log\psi_{2,3}(x_2, x_3|\theta_{2,3}) - \log Z(\theta)\Big\}$$

The vector (x_1, x_2) takes four values $(0,0), (0,1), (1,0), (1,1)$. Take

$$\theta_{1,2} \;:=\; \begin{bmatrix}\log\psi_{1,2}(0,0)\\ \log\psi_{1,2}(0,1)\\ \log\psi_{1,2}(1,0)\\ \log\psi_{1,2}(1,1)\end{bmatrix} \;\in\; \mathbb{R}^4.$$

and let $\psi_{1,2}(x_1, x_2)$ be the function that satisfies

$$\phi_{1,2}(0,0) = \begin{bmatrix}1\\0\\0\\0\end{bmatrix},\;\; \phi_{1,2}(0,1) = \begin{bmatrix}0\\1\\0\\0\end{bmatrix},\;\; \phi_{1,2}(1,0) = \begin{bmatrix}0\\0\\1\\0\end{bmatrix},\;\; \phi_{1,2}(1,1) = \begin{bmatrix}0\\0\\0\\1\end{bmatrix}.$$

With these definitions $\log\psi_{1,2}(x_1, x_2|\theta_{1,2}) = \theta_{1,2}^\top\phi_{1,2}(x_1, x_2)$. We define $\theta_{2,3}$ and $\phi_{2,3}(x_2, x_3)$ in a similar way obtaining that

$$p(x_1, x_2, x_3|\theta) \;=\; \exp\Big\{\theta_{1,2}^\top\phi_{1,2}(x_1, x_2) + \theta_{2,3}^\top\phi_{2,3}(x_2, x_3) - \log Z(\theta)\Big\},$$

which forms an exponential family with sufficient statistics

$$\phi_{1,2}(x_1, x_2) = \begin{bmatrix}(1-x_1)(1-x_2)\\ (1-x_1)x_2\\ x_1(1-x_2)\\ x_1x_2\end{bmatrix}, \qquad \phi_{2,3}(x_2, x_3) = \begin{bmatrix}(1-x_2)(1-x_3)\\ (1-x_2)x_3\\ x_2(1-x_3)\\ x_2x_3\end{bmatrix}$$

and with $Z(\theta) = 1$.

As a **side comment** we note that this exponential family is not minimal in the sense that the values of $\phi_{1,2}(x_1, x_2)$ and $\phi_{2,3}(x_2, x_3)$ lie in a hyperplane in the sense that

$$\phi_{1,2}(x_1,x_2)^\top \begin{bmatrix}1\\1\\1\\1\end{bmatrix} \;=\; 1 \qquad \text{for all } (x_1,x_2) \in \{0,1\}^2.$$

Non-minimal exponential families do not satisfy the gradient equation $\nabla A(\theta) = \mathbb{E}_\theta T(X)$ -- indeed, here $A(\theta) = 0$. An easy solution is to get rid of the first coordinate in $\phi_{1,2}(x_1, x_2)$ and replace it with the corresponding functions of the remaining entries of $\phi_{1,2}(x_1, x_2)$. This defines new natural parameters

$$\bar\theta_{1,2} \;=\; \begin{bmatrix}\log\psi_{1,2}(0,1) - \log\psi_{1,2}(0,0)\\ \log\psi_{1,2}(1,0) - \log\psi_{1,2}(0,0)\\ \log\psi_{1,2}(1,1) - \log\psi_{1,2}(0,0)\end{bmatrix}, \qquad \bar\theta_{2,3} \;=\; \begin{bmatrix}\log\psi_{2,3}(0,1) - \log\psi_{2,3}(0,0)\\ \log\psi_{2,3}(1,0) - \log\psi_{2,3}(0,0)\\ \log\psi_{2,3}(1,1) - \log\psi_{2,3}(0,0)\end{bmatrix}$$

and new sufficient statistics

$$\bar\phi_{1,2}(x_1,x_2) = \begin{bmatrix}(1-x_1)x_2\\ x_1(1-x_2)\\ x_1x_2\end{bmatrix}, \qquad \bar\phi_{2,3}(x_2,x_3) = \begin{bmatrix}(1-x_2)x_3\\ x_2(1-x_3)\\ x_2x_3\end{bmatrix}$$

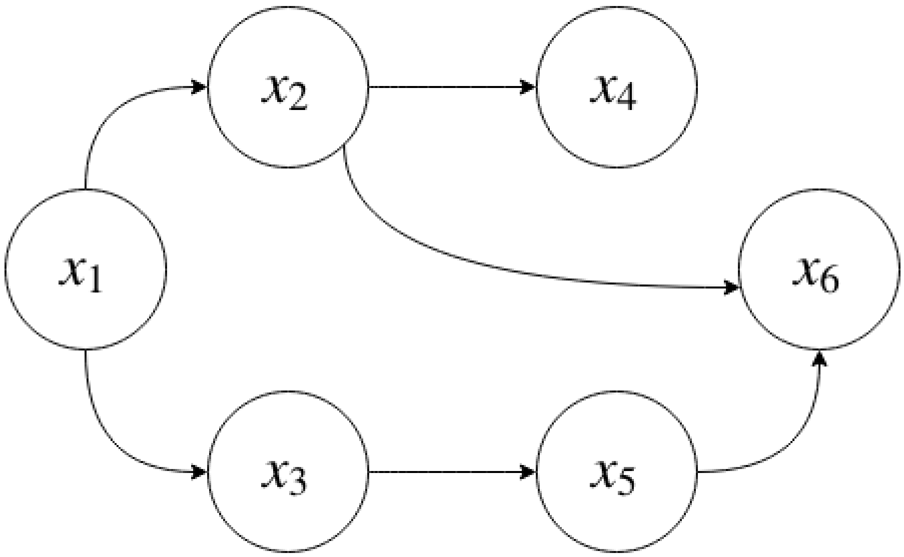
Moreover,

$$A(\bar\theta) \;=\; \log\psi_{1,2}(0,0)\psi_{2,3}(0,0),$$

which should be now be explicitly expressed in terms of $\bar\theta_{1,2}$ and $\bar\theta_{2,3}$.

Simple variable elimination example

Consider the following DAG



Suppose that we observe the variable $X_6 = \bar{x}_6$. What is $p(X_1|\bar{x}_6)$?

The corresponding DAG model implies the factorization:

$$p(x_1, \ldots, x_6) = p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_3)p(x_6|x_2, x_5)$$

We have

$$x_F = \{x_1\}, \qquad x_E = \{x_6\}, \qquad x_R = \{x_2, x_3, x_4, x_5\}$$

$$p(x_F|x_E) = \frac{\sum_{x_R} p(x_F, x_E, x_R)}{\sum_{x_F, x_R} p(x_F, x_E, x_R)}$$

$$\Rightarrow p(x_1|\bar{x}_6) = \frac{p(x_1, \bar{x}_6)}{p(\bar{x}_6)} = \frac{p(x_1, \bar{x}_6)}{\sum_{y_1} p(y_1, \bar{x}_6)}$$

To compute $p(x_1, \bar{x}_6)$, we use variable elimination in the order $2, 3, 4, 5$

$$\begin{aligned} p(x_1, \bar{x}_6) &= p(x_1) \sum_{x_2} \sum_{x_3} \sum_{x_4} \sum_{x_5} p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_3)p(\bar{x}_6|x_2, x_5) \\ &= p(x_1) \sum_{x_2} p(x_2|x_1) \sum_{x_3} p(x_3|x_1) \sum_{x_4} p(x_4|x_2) \sum_{x_5} p(x_5|x_3)p(\bar{x}_6|x_2, x_5) \\ &= p(x_1) \sum_{x_2} p(x_2|x_1) \sum_{x_3} p(x_3|x_1) \sum_{x_4} p(x_4|x_2)p(\bar{x}_6|x_2, x_3) \end{aligned}$$

Note that $p(\bar{x}_6|x_2, x_3)$ does not need to participate in $\sum_{x_4}$.

$$\begin{aligned} &= p(x_1) \sum_{x_2} p(x_2|x_1) \sum_{x_3} p(x_3|x_1)p(\bar{x}_6|x_2, x_3) \sum_{x_4} p(x_4|x_2) \\ &= p(x_1) \sum_{x_2} p(x_2|x_1) \sum_{x_3} p(x_3|x_1)p(\bar{x}_6|x_2, x_3) \\ &= p(x_1) \sum_{x_2} p(x_2|x_1)p(\bar{x}_6|x_1, x_2) \\ &= p(x_1)p(\bar{x}_6|x_1) \end{aligned}$$

Finally,

$$p(x_1|\bar{x}_6) = \frac{p(x_1)p(\bar{x}_6|x_1)}{\sum_{y_1} p(y_1)p(\bar{x}_6|y_1)}.$$

Restricted Boltzmann machines

A restricted Boltzmann machine (RBM) is a simple generative stochastic artificial neural network model. In the language of todays lecture, it is obtained from a special form of the Ising model with variables $(X_1, \ldots, X_k, H_1, \ldots, H_l) \in \{-1, 1\}^{k+1}$. The underlying graph is the bipartite graph with all pairs $H_i - X_j$ connected but with no other edges. Write $\mathbf{x} = (x_1, \ldots, x_k)$, $\mathbf{h} = (h_1, \ldots, h_l)$. The Ising model is then given by all distributions

$$p(\mathbf{x}, \mathbf{h}) \ \propto \ \exp\{\sum_{i=1}^{k} \alpha_i x_i + \sum_{j=1}^{l} \beta_j h_j + \sum_{i=1}^{k} \sum_{j=1}^{l} J_{ij} x_i h_j\},$$

which we can write it in terms of factors

$$\psi_{X_i,H_j}(x_i, h_j) = \exp\{\frac{1}{l}\alpha_i x_i + \frac{1}{k}\beta_j h_j + J_{ij} x_i h_j\}$$

so that

$$p(\mathbf{x}, \mathbf{h}) \ = \ \frac{1}{Z} \prod_{i=1}^{k} \prod_{j=1}^{l} \psi_{X_i,H_j}(x_i, h_j).$$

(Indeed, $\sum_{i=1}^{k} \sum_{j=1}^{l} (\frac{1}{l}\alpha_i x_i + \frac{1}{k}\beta_j h_j + J_{ij} x_i h_j) = \sum_{i=1}^{k} \alpha_i x_i + \sum_{j=1}^{l} \beta_j h_j + \sum_{i=1}^{k} \sum_{j=1}^{l} J_{ij} x_i h_j$)
The normalizing constant $Z = Z(\alpha, \beta, J)$ satisfies

$$Z \ = \ \sum_{\mathbf{x} \in \{-1,1\}^k} \sum_{\mathbf{h} \in \{-1,1\}^l} \prod_{i=1}^{k} \prod_{j=1}^{l} \psi_{X_i,H_j}(x_i, h_j).$$

Note that computing Z may be computationally expensive but we will see that many quantities can be computed without knowing Z. We will need to exploit the structure of the problem.

The corresponding RBM is given as the family of marginal distributions

$$p(\mathbf{x}) = \sum_{\mathbf{h} \in \{-1,1\}^l} p(\mathbf{x}, \mathbf{h}).$$

Denote

$$\tau_j(\mathbf{x}, h_j) \ = \ \prod_{i=1}^{k} \psi_{X_i,H_j}(x_i, h_j) \ = \ \exp\left\{\tfrac{1}{l}\sum_{i=1}^{k} \alpha_i x_i + \beta_j h_j + \sum_{i=1}^{k} J_{ij} x_i h_j\right\},$$

which gives

$$p(\mathbf{x}, \mathbf{h}) \;=\; \frac{1}{Z}\prod_{j=1}^{l}\tau_j(\mathbf{x}, h_j),$$

and note that

$$\begin{aligned}
p(\mathbf{x}) &= \frac{1}{Z}\sum_{\mathbf{h}\in\{-1,1\}^l}\prod_{j=1}^{l}\tau_j(\mathbf{x}, h_j)\\
&= \frac{1}{Z}\left(\sum_{h_1\in\{-1,1\}}\tau_1(\mathbf{x}, h_1)\right)\left(\sum_{h_2\in\{-1,1\}}\tau_2(\mathbf{x}, h_2)\right)\cdots\left(\sum_{h_l\in\{-1,1\}}\tau_l(\mathbf{x}, h_l)\right)\\
&= \frac{1}{Z}\prod_{j=1}^{l}(\tau_j(\mathbf{x}, -1) + \tau_j(\mathbf{x}, 1)).
\end{aligned}$$

We can now easily compute the conditional $p(\mathbf{h}|\mathbf{x})$ and this computation does not even require any knowledge of the normalizing constant Z. For example,

$$\begin{aligned}
p(\mathbf{h}|\mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{h})}{p(\mathbf{x})} &= \frac{\frac{1}{Z}\prod_{j=1}^{l}\tau_j(\mathbf{x}, h_j)}{\frac{1}{Z}\prod_{j=1}^{l}(\tau_j(\mathbf{x}, -1) + \tau_j(\mathbf{x}, 1))}\\
&= \prod_{j=1}^{l}\left(\frac{\tau_j(\mathbf{x}, h_j)}{\tau_j(\mathbf{x}, -1) + \tau_j(\mathbf{x}, 1)}\right).
\end{aligned}$$

We now argue that the bracketed terms above are equal to the conditional probabilities $p(h_j|\mathbf{x})$. Indeed, for example, for $j = 1$ we get

$$p(h_1|\mathbf{x}) \;=\; \sum_{h_2,...,h_l\in\{-1,1\}} p(\mathbf{h}|\mathbf{x}) \;=\; \sum_{h_2,...,h_l\in\{-1,1\}}\prod_{j=1}^{l}\left(\frac{\tau_j(\mathbf{x}, h_j)}{\tau_j(\mathbf{x}, -1) + \tau_j(\mathbf{x}, 1)}\right) \;=\; \frac{\tau_1(\mathbf{x}, h_1)}{\tau_1(\mathbf{x}, -1) + \tau_1(\mathbf{x}, 1)}.$$

In particular, we conclude that $p(\mathbf{h}|\mathbf{x}) = \prod_{j=1}^{l} p(h_j|\mathbf{x})$, which confirms what we know from the Hammersley-Clifford theorem that all H_i's are mutually independent given the vector X. Further, note that

$$\begin{aligned}
p(h_j = 1|x) &= \frac{\prod_{i=1}^{k}\psi_{ij}(x_i, 1)}{\prod_{i=1}^{k}\psi_{ij}(x_i, -1) + \prod_{i=1}^{k}\psi_{ij}(x_i, 1)}\\
&= \frac{\exp\{\frac{1}{l}\sum_{i=1}^{k}\alpha_i x_i + \beta_j + \sum_{i=1}^{k}J_{ij}x_i\}}{\exp\{\frac{1}{l}\sum_{i=1}^{k}\alpha_i x_i - \beta_j - \sum_{i=1}^{k}J_{ij}x_i\} + \exp\{\frac{1}{l}\sum_{i=1}^{k}\alpha_i x_i + \beta_j + \sum_{i=1}^{k}J_{ij}x_i\}}\\
&= \frac{\exp\{\beta_j + \sum_{i=1}^{k}J_{ij}x_i\}}{\exp\{-\beta_j - \sum_{i=1}^{k}J_{ij}x_i\} + \exp\{\beta_j + \sum_{i=1}^{k}J_{ij}x_i\}}\\
&= \sigma\left(\beta_j + \sum_{i=1}^{k}J_{ij}x_i\right)
\end{aligned}$$

with

$$\sigma(y) = \frac{e^y}{e^{-y} + e^y} = \frac{1}{1 + e^{-2y}}$$

called the sigmoid function. Thus, to determine the probability of $H_j = 1$ for each H_j we simply first apply the linear function $\beta + J^\top \mathbf{x}$ to $\mathbf{x}$ (note that the j-th coordinate is precisely $\beta_j + \sum_{i=1}^k J_{ij} x_j$). Then we apply the activation function $\sigma(\cdot)$ coordinate-wise

$$\begin{bmatrix} p(h_1 = 1|\mathbf{x}) \\ p(h_2 = 1|\mathbf{x}) \\ \cdots \\ p(h_l = 1|\mathbf{x}) \end{bmatrix} = \sigma\left(\beta + J^\top \mathbf{x}\right).$$

(sounds familiar?)

MRF Factor product

Given 3 disjoint sets of variables X, Y, Z and factors $\psi_{X,Y}(X, Y)$, $\psi_{Y,Z}(Y, Z)$ the **factor product** is defined as:

$$\psi_{X,Y,Z}(X, Y, Z) = \psi_{X,Y}(X, Y)\psi_{Y,Z}(Y, Z)$$

Take the example below, where we show $\psi_{A,B}(A, B)$, $\psi_{B,C}(B, C)$ and finally, $\psi_{A,B,C}(A, B, C) = \psi_{A,B}(A, B)\psi_{B,C}(B, C)$.

a^1	b^1	0.5
a^1	b^2	0.8
a^2	b^1	0.1
a^2	b^2	0
a^3	b^1	0.3
a^3	b^2	0.9

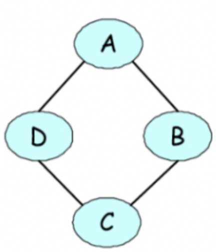
b^1	c^1	0.5
b^1	c^2	0.7
b^2	c^1	0.1
b^2	c^2	0.2

a^1	b^1	c^1	0.5·0.5
a^1	b^1	c^2	0.5·0.7
a^1	b^2	c^1	0.8·0.1
a^1	b^2	c^2	0.8·0.2
a^2	b^1	c^1	0.1·0.5
a^2	b^1	c^2	0.1·0.7
a^2	b^2	c^1	0·0.1
a^2	b^2	c^2	0·0.2
a^3	b^1	c^1	0.3·0.5
a^3	b^1	c^2	0.3·0.7
a^3	b^2	c^1	0.9·0.1
a^3	b^2	c^2	0.9·0.2

Recall our running example from lecture:

$$p(A,B,C,D) = \frac{1}{Z}\psi_{A,B}(A,B)\psi_{B,C}(B,C)\psi_{C,D}(C,D)\psi_{A,D}(A,D)$$

where



$\psi_{AB}[A,B]$			$\psi_{BC}[B,C]$			$\psi_{CD}[C,D]$			$\psi_{AD}[D,A]$		
a^0	b^0	30	b^0	c^0	100	c^0	d^0	1	d^0	a^0	100
a^0	b^1	5	b^0	c^1	1	c^0	d^1	100	d^0	a^1	1
a^1	b^0	1	b^1	c^0	1	c^1	d^0	100	d^1	a^0	1
a^1	b^1	10	b^1	c^1	100	c^1	d^1	1	d^1	a^1	100

From the factor product, we can make queries about the *marginal probabilities*, e.g.

$$\begin{aligned} p(a^0,b^0,c^0,d^0) &\propto \psi_{A,B,C,D}(a^0,b^0,c^0,d^0) \\ &\propto \psi_{A,B}(a^0,b^0)\psi_{B,C}(b^0,c^0)\psi_{C,D}(c^0,d^0)\psi_{A,D}(a^0,d^0) \\ &\propto (30)(100)(1)(100) = 300000 \end{aligned}$$

And if we enumerate all marginal probabilities similarly in a table, we get

Assignment				Unnormalized	Normalized
a^0	b^0	c^0	d^0	300000	0.04
a^0	b^0	c^0	d^1	300000	0.04
a^0	b^0	c^1	d^0	300000	0.04
a^0	b^0	c^1	d^1	30	$4.1 \cdot 10^{-6}$
a^0	b^1	c^0	d^0	500	$6.9 \cdot 10^{-5}$
a^0	b^1	c^0	d^1	500	$6.9 \cdot 10^{-5}$
a^0	b^1	c^1	d^0	5000000	0.69
a^0	b^1	c^1	d^1	500	$6.9 \cdot 10^{-5}$
a^1	b^0	c^0	d^0	100	$1.4 \cdot 10^{-5}$
a^1	b^0	c^0	d^1	1000000	0.14
a^1	b^0	c^1	d^0	100	$1.4 \cdot 10^{-5}$
a^1	b^0	c^1	d^1	100	$1.4 \cdot 10^{-5}$
a^1	b^1	c^0	d^0	10	$1.4 \cdot 10^{-6}$
a^1	b^1	c^0	d^1	100000	0.014
a^1	b^1	c^1	d^0	100000	0.014
a^1	b^1	c^1	d^1	100000	0.014

> To get the normalized marginal probability, divide by the partition function $Z(\theta) = \sum_x \prod_{c \in \mathcal{C}} \psi_c(x_c|\theta_c)$

In order to compute the marginal probability of a single variable in our graph, e.g. $p(b_0)$, marginalize over the other variables:

$$\begin{aligned} p(b^0) &= \sum_{a,c,d} p(a, b^0, c, d) \\ &\propto \sum_{a,c,d} \psi_{A,B,C,D}(a, b^0, c, d) \\ &\propto \sum_{a,c,d} \psi_{A,B}(a, b^0)\psi_{B,C}(b^0, c)\psi_{C,D}(c, d)\psi_{A,D}(a, d) \end{aligned}$$

We can also make queries about the *conditional probability*. Conditioning on an assignment u to a subset of variables U can be done by

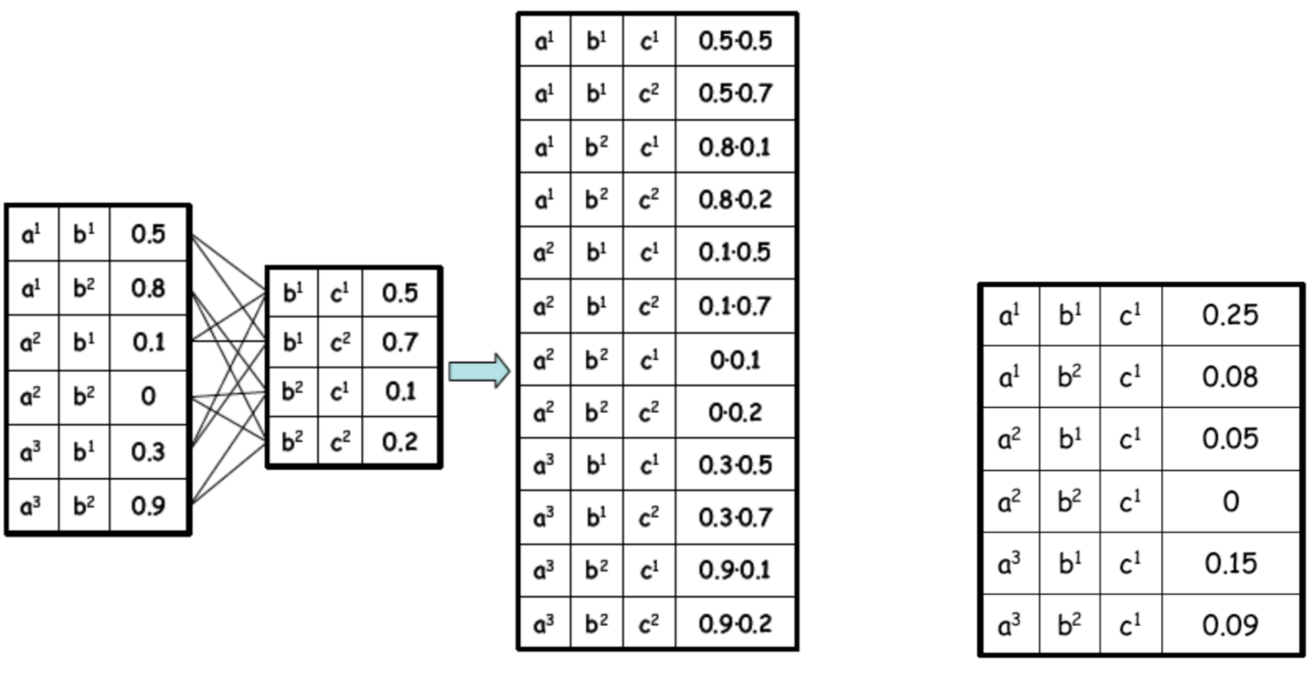
1. Eliminating all entries that are inconsistent with the assignment.
2. Re-normalizing the remaining entries so that they sum to 1.

For example, conditioning on c^1,

$$\begin{aligned} p(a,b|c^1) &= \frac{p(a,b,c^1)}{p(c^1)} \\ &= \frac{p(a,b,c^1)}{\sum_{a,b} p(a,b,c^1)} \\ &= \frac{\psi_{A,B,C}(a,b,c^1)}{\sum_{a,b} \psi_{A,B,C}(a,b,c^1)} \\ &= \frac{\psi_{A,B}(a,b)\psi_{B,C}(b,c^1)}{\sum_{a,b} \psi_{A,B}(a,b)\psi_{B,C}(b,c^1)} \end{aligned}$$

(Note that the original normalization term cancels out in the numerator and denominator.)

Thus, we take only factors consistent with the assignment c^1 and re-normalize with the marginal probability of the variable being conditioned on.



(Original) (Cond. on c^1)

Variable Eliminiation Examples

Example 1:

Take the following factorization:

$$p(C,D,G,H,I,J,L,S) \propto \phi(C)\phi(C,D)\phi(J,L,S)\phi(S,I)\phi(I)\phi(G,D,I)\phi(L,G)\phi(H,G,J)$$

Let's eliminate the variables according to the ordering $\prec \{G,I,S,L,H,C,D\}$.

$$\begin{aligned}
p(J) &= \sum_D \sum_C \phi(C)\phi(C,D) \sum_H \sum_L \sum_S \phi(J,L,S) \sum_I \phi(S,I)\phi(I) \underbrace{\sum_G \phi(G,D,I)\phi(L,G)\phi(H,G,J)}_{\tau(D,I,L,H,J),N_G=6} \\
&= \sum_D \sum_C \phi(C)\phi(C,D) \sum_H \sum_L \sum_S \phi(J,L,S) \underbrace{\sum_I \phi(S,I)\phi(I)\tau(D,I,L,H,J)}_{\tau(D,L,H,J,S),N_I=6} \\
&= \sum_D \sum_C \phi(C)\phi(C,D) \sum_H \sum_L \underbrace{\sum_S \phi(J,L,S)\tau(D,L,H,J,S)}_{\tau(D,L,H,J),N_S=5} \\
&= \sum_D \sum_C \phi(C)\phi(C,D) \sum_H \underbrace{\sum_L \tau(D,L,H,J)}_{\tau(D,H,J),N_L=4} \\
&= \sum_D \sum_C \phi(C)\phi(C,D) \underbrace{\sum_H \tau(D,H,J)}_{\tau(D,J),N_H=3} \\
&= \sum_D \tau(D,J) \underbrace{\sum_C \phi(C)\phi(C,D)}_{\tau(D),N_C=2} \\
&= \underbrace{\sum_D \tau(D,J)\tau(D)}_{\tau(J),N_D=2} \\
&= \tau(J)
\end{aligned}$$

This is a variable elimination ordering over $m = 8$ (initial) factors each with k states.

The sum with the largest number of variables participating has $N_{\max} = 6$ so the complexity is

$$O(8k^6)$$

Note that this is an upper bound.

Example 2:

Let's instead try the Elimination Ordering $\prec \{D,C,H,L,S,I,G\}$,

$$
\begin{aligned}
p(J) &= \sum_G \sum_I \phi(I) \sum_S \phi(S,I) \sum_L \phi(L,G)\phi(J,L,S) \sum_H \phi(H,G,J) \sum_C \phi(C) \underbrace{\sum_D \phi(G,D,I)\phi(C,D)}_{\tau(G,I,C),N_D=4} \\
&= \sum_G \sum_I \phi(I) \sum_S \phi(S,I) \sum_L \phi(L,G)\phi(J,L,S) \sum_H \phi(H,G,J) \underbrace{\sum_C \phi(C)\tau(G,I,C)}_{\tau(G,I),N_C=3} \\
&= \sum_G \sum_I \phi(I)\tau(G,I) \sum_S \phi(S,I) \sum_L \phi(L,G)\phi(J,L,S) \underbrace{\sum_H \phi(H,G,J)}_{\tau(G,J),N_H=3} \\
&= \sum_G \tau(G,J) \sum_I \phi(I)\tau(G,I) \sum_S \phi(S,I) \underbrace{\sum_L \phi(L,G)\phi(J,L,S)}_{\tau(G,J,S),N_L=4} \\
&= \sum_G \tau(G,J) \sum_I \phi(I)\tau(G,I) \underbrace{\sum_S \phi(S,I)\tau(G,J,S)}_{\tau(I,G,J),N_S=4} \\
&= \sum_G \tau(G,J) \underbrace{\sum_I \phi(I)\tau(G,I)\tau(I,G,J)}_{\tau(G,J),N_I=3} \\
&= \underbrace{\sum_G \tau(G,J)\tau(G,J)}_{\tau(J),N_G=2} \\
&= \tau(J)
\end{aligned}
$$

This is a variable elimination ordering over $m = 8$ initial factors each with k states.

The sum with the largest number of variables participating has $N_{\max} = 4$ so the complexity is

$$O(8k^4)$$

Optional Reading

Some questions were asked about whether some algorithm exists for finding the optimal elimination orderings. Although this problem is NP-complete, there are heuristics that can be used. Some discussion of these can be found in Murphy (section 20.3.2), and Daphne Koller's MOOC on PGMs.