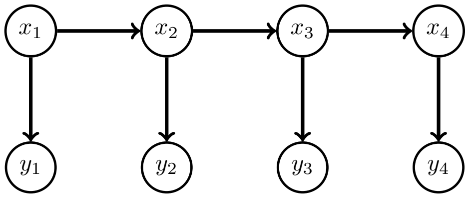
Exercises for the tutorials: 2(a-c) and 4(a-b).

The other exercises are for self-study and exam preparation. All material is examinable unless otherwise mentioned.

Exercise 1. *Conversion to factor graphs*

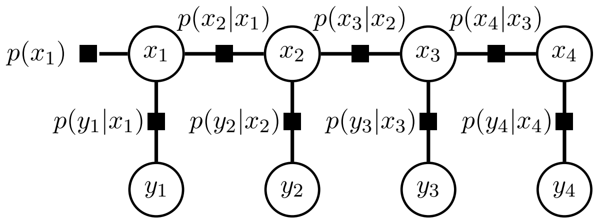
(a) *Draw an undirected factor graph for the directed graphical model defined by the graph below.*



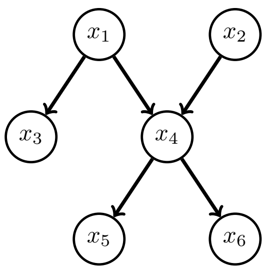
Solution. The graph specifies probabilistic models that factorise as

$$p(x_1, \ldots, x_4, y_1, \ldots, y_4) = p(x_1)p(y_1|x_1) \prod_{i=2}^{4} p(y_i|x_i)p(x_i|x_{i-1})$$

It is the graph for a hidden Markov model. The corresponding factor graph is shown below.



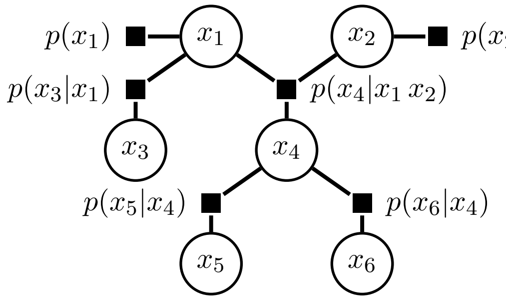
(b) *Draw an undirected factor graph for directed graphical models defined by the graph below (this kind of graph is called a polytree: there are no loops but a node may have more than one parent).*

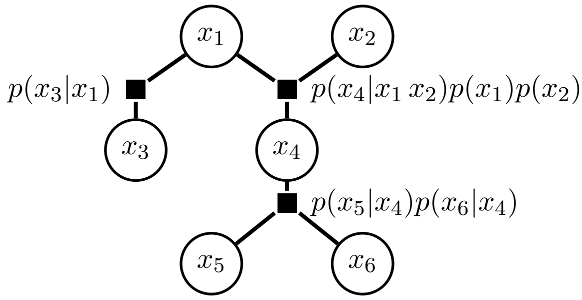


Solution. For the factor graph, we note that the directed graph defines the following class of probabilistic models

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

This gives the factor graph on left in the figure below.



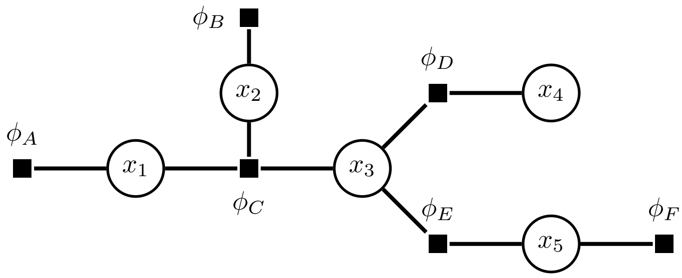


Note:

- One may drop the labels for the factors if the meaning is clear.
- One may choose to group some factors together in order to obtain a factor graph with a particular structure (see factor graph on right)

Exercise 2. ***Sum-product message passing***

We here re-consider the factor tree from the lecture on exact inference.



Let all variables be binary, $x_i \in \{0, 1\}$, and the factors be defined as follows:

x_1	ϕ_A
0	2
1	4

x_2	ϕ_B
0	4
1	4

x_1	x_2	x_3	ϕ_C
0	0	0	4
1	0	0	2
0	1	0	2
1	1	0	6
0	0	1	2
1	0	1	6
0	1	1	6
1	1	1	4

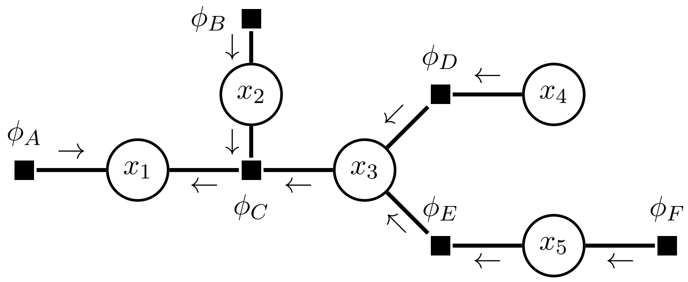
x_3	x_4	ϕ_D
0	0	8
1	0	2
0	1	2
1	1	6

x_3	x_5	ϕ_E
0	0	3
1	0	6
0	1	6
1	1	3

x_5	ϕ_F
0	1
1	8

(a) *Mark the graph with arrows indicating all messages that need to be computed for the computation of $p(x_1)$.*

Solution.



(b) *Compute the messages that you have identified.*

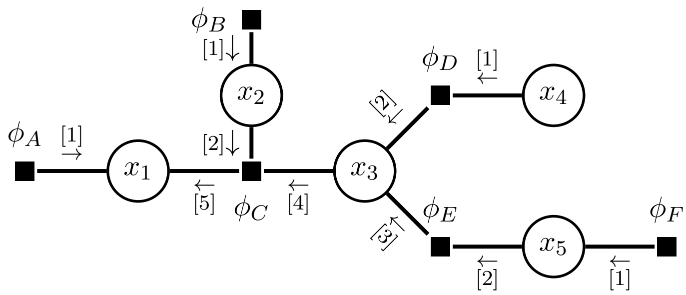
Assuming that the computation of the messages is scheduled according to a common clock, group the messages together so that all messages in the same group can be computed in parallel during a clock cycle.

Solution. Since the variables are binary, each message can be represented as a two-dimensional vector. We use the convention that the first element of the vector corresponds to the message for $x_i = 0$ and the second element to the message for $x_i = 1$. For example,

$$\boldsymbol{\mu_{\phi_A \to x_1}} = \begin{pmatrix} 2 \\ 4 \end{pmatrix} \tag{S.1}$$

means that the message $\mu_{\phi_A \to x_1}(x_1)$ equals 2 for $x_1 = 0$, i.e. $\mu_{\phi_A \to x_1}(0) = 2$.

The following figure shows a grouping (scheduling) of the computation of the messages.



Clock cycle 1:

$$\boldsymbol{\mu_{\phi_A \to x_1}} = \begin{pmatrix} 2 \\ 4 \end{pmatrix} \qquad \boldsymbol{\mu_{\phi_B \to x_2}} = \begin{pmatrix} 4 \\ 4 \end{pmatrix} \qquad \boldsymbol{\mu_{x_4 \to \phi_D}} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \qquad \boldsymbol{\mu_{\phi_F \to x_5}} = \begin{pmatrix} 1 \\ 8 \end{pmatrix} \tag{S.2}$$

Clock cycle 2:

$$\boldsymbol{\mu_{x_2 \to \phi_C}} = \boldsymbol{\mu_{\phi_B \to x_2}} = \begin{pmatrix} 4 \\ 4 \end{pmatrix} \qquad \boldsymbol{\mu_{x_5 \to \phi_E}} = \boldsymbol{\mu_{\phi_F \to x_5}} = \begin{pmatrix} 1 \\ 8 \end{pmatrix} \tag{S.3}$$

Message $\mu_{\phi_D \to x_3}$ is defined as

$$\mu_{\phi_D \to x_3}(x_3) = \sum_{x_4} \phi_D(x_3, x_4)\mu_{x_4 \to \phi_D}(x_4) \tag{S.4}$$

so that

$$\mu_{\phi_D \to x_3}(0) = \sum_{x_4=0}^{1} \phi_D(0, x_4)\mu_{x_4 \to \phi_D}(x_4) \tag{S.5}$$
$$= \phi_D(0,0)\mu_{x_4 \to \phi_D}(0) + \phi_D(0,1)\mu_{x_4 \to \phi_D}(1) \tag{S.6}$$
$$= 8 \cdot 1 + 2 \cdot 1 \tag{S.7}$$
$$= 10 \tag{S.8}$$
$$\mu_{\phi_D \to x_3}(1) = \sum_{x_4=0}^{1} \phi_D(1, x_4)\mu_{x_4 \to \phi_D}(x_4) \tag{S.9}$$
$$= \phi_D(1,0)\mu_{x_4 \to \phi_D}(0) + \phi_D(1,1)\mu_{x_4 \to \phi_D}(1) \tag{S.10}$$
$$= 2 \cdot 1 + 6 \cdot 1 \tag{S.11}$$
$$= 8 \tag{S.12}$$

and thus

$$\boldsymbol{\mu_{\phi_D \to x_3}} = \begin{pmatrix} 10 \\ 8 \end{pmatrix}. \tag{S.13}$$

The above computations can be written more compactly in matrix notation. Let $\boldsymbol{\phi_D}$ be the matrix that contains the outputs of $\phi_D(x_3, x_4)$

$$\boldsymbol{\phi_D} = \begin{pmatrix} \phi_D(x_3 = 0, x_4 = 0) & \phi_D(x_3 = 0, x_4 = 1) \\ \phi_D(x_3 = 1, x_4 = 0) & \phi_D(x_3 = 1, x_4 = 1) \end{pmatrix} = \begin{pmatrix} 8 & 2 \\ 2 & 6 \end{pmatrix}. \tag{S.14}$$

We can then write $\boldsymbol{\mu_{\phi_D \to x_3}}$ in terms of a matrix vector product,

$$\boldsymbol{\mu_{\phi_D \to x_3}} = \boldsymbol{\phi_D} \boldsymbol{\mu_{x_4 \to \phi_D}}. \tag{S.15}$$

Clock cycle 3:
Representing the factor ϕ_E as matrix $\boldsymbol{\phi_E}$,

$$\boldsymbol{\phi_E} = \begin{pmatrix} \phi_E(x_3 = 0, x_5 = 0) & \phi_E(x_3 = 0, x_5 = 1) \\ \phi_E(x_3 = 1, x_5 = 0) & \phi_E(x_3 = 1, x_5 = 1) \end{pmatrix} = \begin{pmatrix} 3 & 6 \\ 6 & 3 \end{pmatrix}, \tag{S.16}$$

we can write

$$\mu_{\phi_E \to x_3}(x_3) = \sum_{x_5} \phi_E(x_3, x_5)\mu_{x_5 \to \phi_E}(x_5) \tag{S.17}$$

as a matrix vector product,

$$\boldsymbol{\mu_{\phi_E \to x_3}} = \boldsymbol{\phi_E} \boldsymbol{\mu_{x_5 \to \phi_E}} \tag{S.18}$$
$$= \begin{pmatrix} 3 & 6 \\ 6 & 3 \end{pmatrix} \begin{pmatrix} 1 \\ 8 \end{pmatrix} \tag{S.19}$$
$$= \begin{pmatrix} 51 \\ 30 \end{pmatrix}. \tag{S.20}$$

Clock cycle 4:
Variable node x_3 has received all incoming messages, and can thus output $\mu_{x_3 \to \phi_C}$,

$$\mu_{x_3 \to \phi_C}(x_3) = \mu_{\phi_D \to x_3}(x_3)\mu_{\phi_E \to x_3}(x_3). \tag{S.21}$$

Using $\odot$ to denote element-wise multiplication of two vectors, we have

$$\boldsymbol{\mu}_{\boldsymbol{x_3}\to\boldsymbol{\phi_C}} = \boldsymbol{\mu}_{\boldsymbol{\phi_D}\to\boldsymbol{x_3}} \odot \boldsymbol{\mu}_{\boldsymbol{\phi_E}\to\boldsymbol{x_3}} \tag{S.22}$$
$$= \begin{pmatrix}10\\8\end{pmatrix} \odot \begin{pmatrix}51\\30\end{pmatrix} \tag{S.23}$$
$$= \begin{pmatrix}510\\240\end{pmatrix}. \tag{S.24}$$

Clock cycle 5:
Factor node ϕ_C has received all incoming messages, and can thus output $\mu_{\phi_C\to x_1}$,

$$\mu_{\phi_C\to x_1}(x_1) = \sum_{x_2,x_3} \phi_C(x_1,x_2,x_3)\mu_{x_2\to\phi_C}(x_2)\mu_{x_3\to\phi_C}(x_3). \tag{S.25}$$

Writing out the sum for $x_1 = 0$ and $x_1 = 1$ gives

$$\mu_{\phi_C\to x_1}(0) = \sum_{x_2,x_3} \phi_C(0,x_2,x_3)\mu_{x_2\to\phi_C}(x_2)\mu_{x_3\to\phi_C}(x_3) \tag{S.26}$$
$$=\phi_C(0,x_2,x_3)\mu_{x_2\to\phi_C}(x_2)\mu_{x_3\to\phi_C}(x_3)\ |_{(x_2,x_3)=(0,0)} + \tag{S.27}$$
$$\phi_C(0,x_2,x_3)\mu_{x_2\to\phi_C}(x_2)\mu_{x_3\to\phi_C}(x_3)\ |_{(x_2,x_3)=(1,0)} + \tag{S.28}$$
$$\phi_C(0,x_2,x_3)\mu_{x_2\to\phi_C}(x_2)\mu_{x_3\to\phi_C}(x_3)\ |_{(x_2,x_3)=(0,1)} + \tag{S.29}$$
$$\phi_C(0,x_2,x_3)\mu_{x_2\to\phi_C}(x_2)\mu_{x_3\to\phi_C}(x_3)\ |_{(x_2,x_3)=(1,1)} \tag{S.30}$$
$$=4\cdot 4\cdot 510+ \tag{S.31}$$
$$2\cdot 4\cdot 510+ \tag{S.32}$$
$$2\cdot 4\cdot 240+ \tag{S.33}$$
$$6\cdot 4\cdot 240 \tag{S.34}$$
$$=19920 \tag{S.35}$$
$$\mu_{\phi_C\to x_1}(1) = \sum_{x_2,x_3} \phi_C(1,x_2,x_3)\mu_{x_2\to\phi_C}(x_2)\mu_{x_3\to\phi_C}(x_3) \tag{S.36}$$
$$=\phi_C(1,x_2,x_3)\mu_{x_2\to\phi_C}(x_2)\mu_{x_3\to\phi_C}(x_3)\ |_{(x_2,x_3)=(0,0)} + \tag{S.37}$$
$$\phi_C(1,x_2,x_3)\mu_{x_2\to\phi_C}(x_2)\mu_{x_3\to\phi_C}(x_3)\ |_{(x_2,x_3)=(1,0)} + \tag{S.38}$$
$$\phi_C(1,x_2,x_3)\mu_{x_2\to\phi_C}(x_2)\mu_{x_3\to\phi_C}(x_3)\ |_{(x_2,x_3)=(0,1)} + \tag{S.39}$$
$$\phi_C(1,x_2,x_3)\mu_{x_2\to\phi_C}(x_2)\mu_{x_3\to\phi_C}(x_3)\ |_{(x_2,x_3)=(1,1)} \tag{S.40}$$
$$=2\cdot 4\cdot 510+ \tag{S.41}$$
$$6\cdot 4\cdot 510+ \tag{S.42}$$
$$6\cdot 4\cdot 240+ \tag{S.43}$$
$$4\cdot 4\cdot 240 \tag{S.44}$$
$$=25920 \tag{S.45}$$

and hence

$$\boldsymbol{\mu}_{\boldsymbol{\phi_C}\to\boldsymbol{x_1}} = \begin{pmatrix}19920\\25920\end{pmatrix} \tag{S.46}$$

After step 5, variable node x_1 has received all incoming messages and the marginal can be computed.

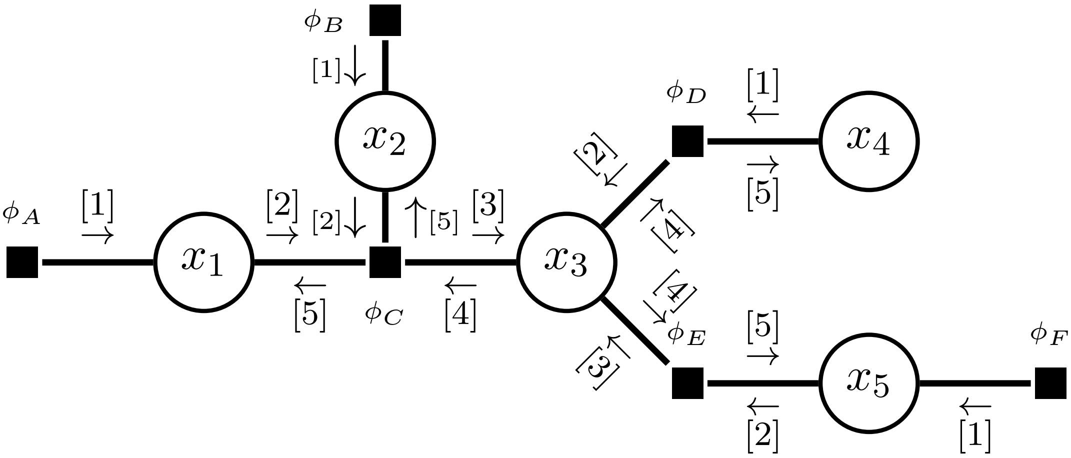


Figure 1: Answer to Exercise 2 Question (b): Computing all messages in five clock cycles. If we also computed the messages toward the leaf factor nodes, we needed six cycles, but they are not necessary for computation of the marginals so they are omitted.

In addition to the messages needed for computation of $p(x_1)$ one can compute *all* messages in the graph in five clock cycles, see Figure 1. This means that *all* marginals, as well as the joints of those variables sharing a factor node, are available after five clock cycles.

(c) *What is $p(x_1 = 1)$?*

Solution. We compute the marginal $p(x_1)$ as

$$p(x_1) \propto \mu_{\phi_A \to x_1}(x_1)\mu_{\phi_C \to x_1}(x_1) \tag{S.47}$$

which is in vector notation

$$\begin{pmatrix} p(x_1 = 0) \\ p(x_1 = 1) \end{pmatrix} \propto \boldsymbol{\mu}_{\boldsymbol{\phi_A \to x_1}} \odot \boldsymbol{\mu}_{\boldsymbol{\phi_C \to x_1}} \tag{S.48}$$

$$\propto \begin{pmatrix} 2 \\ 4 \end{pmatrix} \odot \begin{pmatrix} 19920 \\ 25920 \end{pmatrix} \tag{S.49}$$

$$\propto \begin{pmatrix} 39840 \\ 103680 \end{pmatrix}. \tag{S.50}$$

Normalisation gives

$$\begin{pmatrix} p(x_1 = 0) \\ p(x_1 = 1) \end{pmatrix} = \frac{1}{39840 + 103680} \begin{pmatrix} 39840 \\ 103680 \end{pmatrix} \tag{S.51}$$

$$= \begin{pmatrix} 0.2776 \\ 0.7224 \end{pmatrix} \tag{S.52}$$

so that $p(x_1 = 1) = 0.7224$.

Note the relatively large numbers in the messages that we computed. In other cases, one may obtain very small ones depending on the scale of the factors. This can cause numerical issues that can be addressed by working in the logarithmic domain.

(d) *Draw the factor graph corresponding to $p(x_1, x_3, x_4, x_5 | x_2 = 1)$ and provide the numerical values for all factors.*

Solution. The pmf represented by the original factor graph is

$$p(x_1, \ldots, x_5) \propto \phi_A(x_1)\phi_B(x_2)\phi_C(x_1, x_2, x_3)\phi_D(x_3, x_4)\phi_E(x_3, x_5)\phi_F(x_5)$$

The conditional $p(x_1, x_3, x_4, x_5|x_2 = 1)$ is proportional to $p(x_1, \ldots, x_5)$ with x_2 fixed to $x_2 = 1$, i.e.

$$p(x_1, x_3, x_4, x_5|x_2 = 1) \propto p(x_1, x_2 = 1, x_3, x_4, x_5) \tag{S.53}$$

$$\propto \phi_A(x_1)\phi_B(x_2 = 1)\phi_C(x_1, x_2 = 1, x_3)\phi_D(x_3, x_4)\phi_E(x_3, x_5)\phi_F(x_5) \tag{S.54}$$

$$\propto \phi_A(x_1)\phi_C^{x_2}(x_1, x_3)\phi_D(x_3, x_4)\phi_E(x_3, x_5)\phi_F(x_5) \tag{S.55}$$

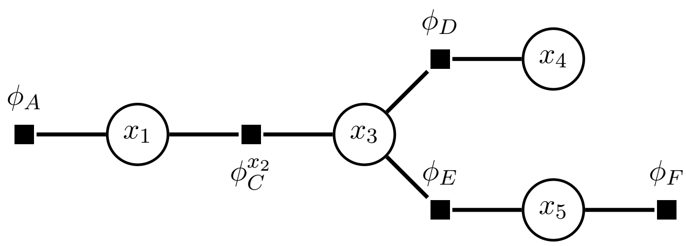
where $\phi_C^{x_2}(x_1, x_3) = \phi_C(x_1, x_2 = 1, x_3)$. The numerical values of $\phi_C^{x_2}(x_1, x_3)$ can be read from the table defining $\phi_C(x_1, x_2, x_3)$, extracting those rows where $x_2 = 1$,

	x_1	x_2	x_3	ϕ_C
	0	0	0	4
	1	0	0	2
$\rightarrow$	0	1	0	2
$\rightarrow$	1	1	0	6
	0	0	1	2
	1	0	1	6
$\rightarrow$	0	1	1	6
$\rightarrow$	1	1	1	4

so that

x_1	x_3	$\phi_C^{x_2}$
0	0	2
1	0	6
0	1	6
1	1	4

The factor graph for $p(x_1, x_3, x_4, x_5|x_2 = 1)$ is shown below. Factor ϕ_B has disappeared since it only depended on x_2 and thus became a constant. Factor ϕ_C is replaced by $\phi_C^{x_2}$ defined above. The remaining factors are the same as in the original factor graph.



(e) *Compute $p(x_1 = 1|x_2 = 1)$, re-using messages that you have already computed for the evaluation of $p(x_1 = 1)$.*

Solution. The message $\mu_{\phi_A \to x_1}$ is the same as in the original factor graph and $\mu_{x_3 \to \phi_C^{x_2}} = \mu_{x_3 \to \phi_C}$. This is because the outgoing message from x_3 corresponds to the effective factor obtained by summing out all variables in the sub-trees attached to x_3 (without the $\phi_C^{x_2}$ branch), and these sub-trees do not depend on x_2.

The message $\mu_{\phi_C^{x_2} \to x_1}$ needs to be newly computed. We have

$$\mu_{\phi_C^{x_2} \to x_1}(x_1) = \sum_{x_3} \phi_C^{x_2}(x_1, x_3)\mu_{x_3 \to \phi_C^{x_2}} \tag{S.56}$$

or in vector notation

$$\boldsymbol{\mu}_{\phi_C^{x_2}\to x_1} = \boldsymbol{\phi}_C^{x_2}\boldsymbol{\mu}_{x_3\to\phi_C^{x_2}} \tag{S.57}$$

$$= \begin{pmatrix} \phi_C^{x_2}(x_1=0,x_3=0) & \phi_C^{x_2}(x_1=0,x_3=1) \\ \phi_C^{x_2}(x_1=1,x_3=0) & \phi_C^{x_2}(x_1=1,x_3=1) \end{pmatrix} \boldsymbol{\mu}_{x_3\to\phi_C^{x_2}} \tag{S.58}$$

$$= \begin{pmatrix} 2 & 6 \\ 6 & 4 \end{pmatrix} \begin{pmatrix} 510 \\ 240 \end{pmatrix} \tag{S.59}$$

$$= \begin{pmatrix} 2460 \\ 4020 \end{pmatrix} \tag{S.60}$$

We thus obtain for the marginal posterior of x_1 given $x_2 = 1$:

$$\begin{pmatrix} p(x_1=0|x_2=1) \\ p(x_1=1|x_2=1) \end{pmatrix} \propto \boldsymbol{\mu}_{\phi_A\to x_1} \odot \boldsymbol{\mu}_{\phi_C^{x_2}\to x_1} \tag{S.61}$$

$$\propto \begin{pmatrix} 2 \\ 4 \end{pmatrix} \odot \begin{pmatrix} 2460 \\ 4020 \end{pmatrix} \tag{S.62}$$

$$\propto \begin{pmatrix} 4920 \\ 16080 \end{pmatrix}. \tag{S.63}$$

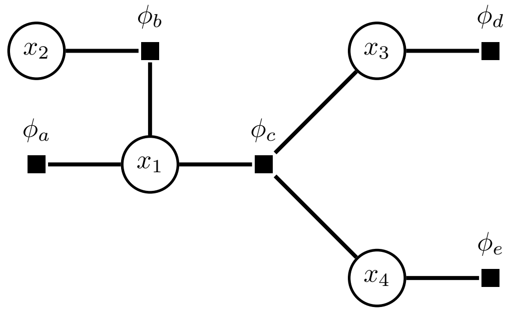
Normalisation gives

$$\begin{pmatrix} p(x_1=0|x_2=1) \\ p(x_1=1|x_2=1) \end{pmatrix} = \begin{pmatrix} 0.2343 \\ 0.7657 \end{pmatrix} \tag{S.64}$$

and thus $p(x_1 = 1|x_2 = 1) = 0.7657$. The posterior probability is slightly larger than the prior probability, $p(x_1 = 1) = 0.7224$.

Exercise 3. *Sum-product message passing*

The following factor graph represents a Gibbs distribution over four binary variables $x_i \in \{0, 1\}$.



The factors ϕ_a, ϕ_b, ϕ_d are defined as follows:

x_1	ϕ_a
0	2
1	1

x_1	x_2	ϕ_b
0	0	5
1	0	2
0	1	2
1	1	6

x_3	ϕ_d
0	1
1	2

and $\phi_c(x_1, x_3, x_4) = 1$ if $x_1 = x_3 = x_4$, and is zero otherwise.

For all questions below, justify your answer:

(a) *Compute the values of $\mu_{x_2 \to \phi_b}(x_2)$ for $x_2 = 0$ and $x_2 = 1$.*

Solution. Messages from leaf-variable nodes to factor nodes are equal to one, so that $\mu_{x_2 \to \phi_b}(x_2) = 1$ for all x_2.

(b) *Assume the message $\mu_{x_4 \to \phi_c}(x_4)$ equals*

$$\mu_{x_4 \to \phi_c}(x_4) = \begin{cases} 1 & \text{if } x_4 = 0 \\ 3 & \text{if } x_4 = 1 \end{cases}$$

Compute the values of $\phi_e(x_4)$ for $x_4 = 0$ and $x_4 = 1$.

Solution. Messages from leaf-factors to their variable nodes are equal to the leaf-factors, and variable nodes with single incoming messages copy the message. We thus have

$$\mu_{\phi_e \to x_4}(x_4) = \phi_e(x_4) \tag{S.65}$$

$$\mu_{x_4 \to \phi_c}(x_4) = \mu_{\phi_e \to x_4}(x_4) \tag{S.66}$$

and hence

$$\phi_e(x_4) = \begin{cases} 1 & \text{if } x_4 = 0 \\ 3 & \text{if } x_4 = 1 \end{cases} \tag{S.67}$$

(c) *Compute the values of $\mu_{\phi_c \to x_1}(x_1)$ for $x_1 = 0$ and $x_1 = 1$.*

Solution. We first compute $\mu_{x_3 \to \phi_c}(x_3)$:

$$\mu_{x_3 \to \phi_c}(x_3) = \mu_{\phi_d \to x_3}(x_3) \tag{S.68}$$

$$= \begin{cases} 1 & \text{if } x_3 = 0 \\ 2 & \text{if } x_3 = 1 \end{cases} \tag{S.69}$$

The desired message $\mu_{\phi_c \to x_1}(x_1)$ is by definition

$$\mu_{\phi_c \to x_1}(x_1) = \sum_{x_3, x_4} \phi_c(x_1, x_3, x_4) \mu_{x_3 \to \phi_c}(x_3) \mu_{x_4 \to \phi_c}(x_4) \tag{S.70}$$

Since $\phi_c(x_1, x_3, x_4)$ is only non-zero if $x_1 = x_3 = x_4$, where it equals one, the computations simplify:

$$\mu_{\phi_c \to x_1}(x_1 = 0) = \phi_c(0, 0, 0) \mu_{x_3 \to \phi_c}(0) \mu_{x_4 \to \phi_c}(0) \tag{S.71}$$

$$= 1 \cdot 1 \cdot 1 \tag{S.72}$$

$$= 1 \tag{S.73}$$

$$\mu_{\phi_c \to x_1}(x_1 = 1) = \phi_c(1, 1, 1) \mu_{x_3 \to \phi_c}(1) \mu_{x_4 \to \phi_c}(1) \tag{S.74}$$

$$= 1 \cdot 2 \cdot 3 \tag{S.75}$$

$$= 6 \tag{S.76}$$

(d) The message $\mu_{\phi_b \to x_1}(x_1)$ equals

$$\mu_{\phi_b \to x_1}(x_1) = \begin{cases} 7 & \text{if } x_1 = 0 \\ 8 & \text{if } x_1 = 1 \end{cases}$$

What is the probability that $x_1 = 1$, i.e. $p(x_1 = 1)$?

Solution. The unnormalised marginal $p(x_1)$ is given by the product of the three incoming messages

$$p(x_1) \propto \mu_{\phi_a \to x_1}(x_1)\mu_{\phi_b \to x_1}(x_1)\mu_{\phi_c \to x_1}(x_1) \tag{S.77}$$

With

$$\mu_{\phi_b \to x_1}(x_1) = \sum_{x_2} \phi_b(x_1, x_2) \tag{S.78}$$

it follows that

$$\begin{aligned} \mu_{\phi_b \to x_1}(x_1 = 0) &= \sum_{x_2} \phi_b(0, x_2) && \text{(S.79)} \\ &= 5 + 2 && \text{(S.80)} \\ &= 7 && \text{(S.81)} \\ \mu_{\phi_b \to x_1}(x_1 = 1) &= \sum_{x_2} \phi_b(1, x_2) && \text{(S.82)} \\ &= 2 + 6 && \text{(S.83)} \\ &= 8 && \text{(S.84)} \end{aligned}$$

Hence, we obtain

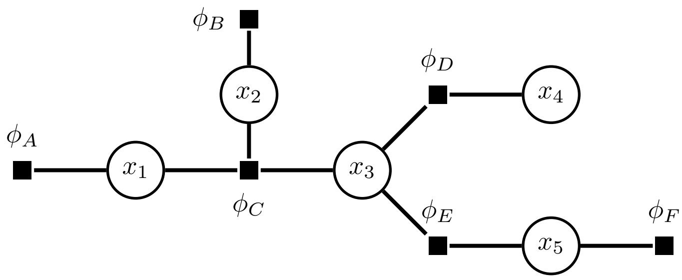
$$\begin{aligned} p(x_1 = 0) &\propto 2 \cdot 7 \cdot 1 = 14 && \text{(S.85)} \\ p(x_1 = 1) &\propto 1 \cdot 8 \cdot 6 = 48 && \text{(S.86)} \end{aligned}$$

and normalisation yields the desired result

$$p(x_1 = 1) = \frac{48}{14 + 48} = \frac{48}{62} = \frac{24}{31} = 0.774 \tag{S.87}$$

Exercise 4. *Max-sum message passing*

We here compute most probable states for the factor graph and factors below.



Let all variables be binary, $x_i \in \{0, 1\}$, and the factors be defined as follows:

x_1	ϕ_A
0	2
1	4

x_2	ϕ_B
0	4
1	4

x_1	x_2	x_3	ϕ_C
0	0	0	4
1	0	0	2
0	1	0	2
1	1	0	6
0	0	1	2
1	0	1	6
0	1	1	6
1	1	1	4

x_3	x_4	ϕ_D
0	0	8
1	0	2
0	1	2
1	1	6

x_3	x_5	ϕ_E
0	0	3
1	0	6
0	1	6
1	1	3

x_5	ϕ_F
0	1
1	8

(a) *Will we need to compute the normalising constant Z to determine* $\operatorname{argmax}_{\mathbf{x}} p(x_1, \ldots, x_5)$*?*

Solution. This is not necessary since $\operatorname{argmax}_{\mathbf{x}} p(x_1, \ldots, x_5) = \operatorname{argmax}_{\mathbf{x}} cp(x_1, \ldots, x_5)$ for any constant c. Algorithmically, the backtracking algorithm is also invariant to any scaling of the factors.

(b) *Compute* $\operatorname{argmax}_{x_1,x_2,x_3} p(x_1, x_2, x_3|x_4 = 0, x_5 = 0)$ *via max-sum message passing.*

Solution. We first derive the factor graph and corresponding factors for $p(x_1, x_2, x_3|x_4 = 0, x_5 = 0)$.

For fixed values of x_4, x_5, the two variables are removed from the graph, and the factors $\phi_D(x_3, x_4)$ and $\phi_E(x_3, x_5)$ are reduced to univariate factors $\phi_D^{x_4}(x_3)$ and $\phi_D^{x_5}(x_3)$ by retaining those rows in the table where $x_4 = 0$ and $x_5 = 0$, respectively:

x_3	$\phi_D^{x_4}$
0	8
1	2

x_3	$\phi_E^{x_5}$
0	3
1	6

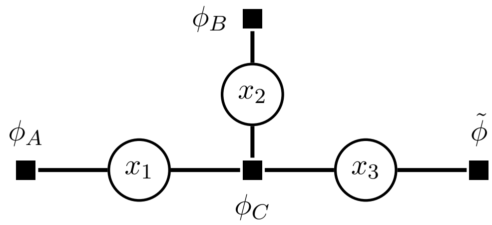
Since both factors only depend on x_3, they can be combined into a new factor $\tilde{\phi}(x_3)$ by element-wise multiplication.

x_3	$\tilde{\phi}$
0	24
1	12

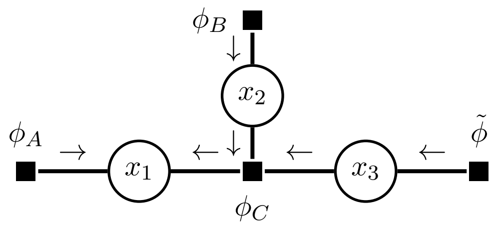
Moreover, since we work with an unnormalised model, we can rescale the factor so that the maximum value is one, so that

x_3	$\tilde{\phi}$
0	2
1	1

Factor $\phi_F(x_5)$ is a constant for fixed value of x_5 and can be ignored. The factor graph for $p(x_1, x_2, x_3|x_4 = 0, x_5 = 0)$ thus is



Let us fix x_1 as root towards which we compute the messages. The messages that we need to compute are shown in the following graph



Next, we compute the leaf (log) messages. We only have factor nodes as leaf nodes so that

$$\boldsymbol{\lambda}_{\phi_A \to x_1} = \begin{pmatrix} \log \phi_A(x_1 = 0) \\ \log \phi_A(x_1 = 1) \end{pmatrix} = \begin{pmatrix} \log 2 \\ \log 4 \end{pmatrix} \tag{S.88}$$

and similarly

$$\boldsymbol{\lambda}_{\phi_B \to x_2} = \begin{pmatrix} \log \phi_B(x_2 = 0) \\ \log \phi_B(x_2 = 1) \end{pmatrix} = \begin{pmatrix} \log 4 \\ \log 4 \end{pmatrix} \quad \boldsymbol{\lambda}_{\tilde{\phi} \to x_3} = \begin{pmatrix} \log \tilde{\phi}(x_3 = 0) \\ \log \tilde{\phi}(x_3 = 1) \end{pmatrix} = \begin{pmatrix} \log 2 \\ \log 1 \end{pmatrix} \tag{S.89}$$

Since the variable nodes x_2 and x_3 only have one incoming edge each, we obtain

$$\boldsymbol{\lambda}_{x_2 \to \phi_C} = \boldsymbol{\lambda}_{\phi_B \to x_2} = \begin{pmatrix} \log 4 \\ \log 4 \end{pmatrix} \qquad \boldsymbol{\lambda}_{x_3 \to \phi_C} = \boldsymbol{\lambda}_{\tilde{\phi} \to x_3} = \begin{pmatrix} \log 2 \\ \log 1 \end{pmatrix} \tag{S.90}$$

The message $\lambda_{\phi_C \to x_1}(x_1)$ equals

$$\lambda_{\phi_C \to x_1}(x_1) = \max_{x_2, x_3} \log \phi_C(x_1, x_2, x_3) + \lambda_{x_2 \to \phi_C}(x_2) + \lambda_{x_3 \to \phi_C}(x_3) \tag{S.91}$$

where we wrote the messages in non-vector notation to highlight their dependency on the variables x_2 and x_3. We now have to consider all combinations of x_2 and x_3

x_2	x_3	$\log \phi_C(x_1 = 0, x_2, x_3)$
0	0	$\log 4$
1	0	$\log 2$
0	1	$\log 2$
1	1	$\log 6$

x_2	x_3	$\log \phi_C(x_1 = 1, x_2, x_3)$
0	0	$\log 2$
1	0	$\log 6$
0	1	$\log 6$
1	1	$\log 4$

Furthermore

x_2	x_3	$\lambda_{x_2 \to \phi_C}(x_2) + \lambda_{x_3 \to \phi_C}(x_3)$
0	0	$\log 4 + \log 2 = \log 8$
1	0	$\log 4 + \log 2 = \log 8$
0	1	$\log 4$
1	1	$\log 4$

Hence for $x_1 = 0$, we have

x_2	x_3	$\log \phi_C(x_1 = 0, x_2, x_3) + \lambda_{x_2 \to \phi_C}(x_2) + \lambda_{x_3 \to \phi_C}(x_3)$
0	0	$\log 4 + \log 8 = \log 32$
1	0	$\log 2 + \log 8 = \log 16$
0	1	$\log 2 + \log 4 = \log 8$
1	1	$\log 6 + \log 4 = \log 24$

The maximal value is $\log 32$ and for backtracking, we also need to keep track of the argmax which is here $\hat{x}_2 = \hat{x}_3 = 0$.

For $x_1 = 1$, we have

x_2	x_3	$\log \phi_C(x_1 = 1, x_2, x_3) + \lambda_{x_2 \to \phi_C}(x_2) + \lambda_{x_3 \to \phi_C}(x_3)$
0	0	$\log 2 + \log 8 = \log 16$
1	0	$\log 6 + \log 8 = \log 48$
0	1	$\log 6 + \log 4 = \log 24$
1	1	$\log 4 + \log 4 = \log 16$

The maximal value is $\log 48$ and the argmax is $(\hat{x}_2 = 1, \hat{x}_3 = 0)$.

So overall, we have

$$\boldsymbol{\lambda}_{\phi_C \to x_1} = \begin{pmatrix} \lambda_{\phi_C \to x_1}(x_1 = 0) \\ \lambda_{\phi_C \to x_1}(x_1 = 1) \end{pmatrix} = \begin{pmatrix} \log 32 \\ \log 48 \end{pmatrix} \tag{S.92}$$

and the argmax back-tracking function is

$$\lambda^*_{\phi_C \to x_1}(x_1) = \begin{cases} (\hat{x}_2 = 0, \hat{x}_3 = 0) & \text{if } x_1 = 0 \\ (\hat{x}_2 = 1, \hat{x}_3 = 0) & \text{if } x_1 = 1 \end{cases} \tag{S.93}$$

We now have all incoming messages to the assigned root node x_1. *Ignoring the normalising constant*, we obtain

$$\boldsymbol{\gamma} = \begin{pmatrix} \gamma^*(x_1 = 0) \\ \gamma^*(x_1 = 1) \end{pmatrix} = \boldsymbol{\lambda}_{\phi_A \to x_1} + \boldsymbol{\lambda}_{\phi_C \to x_1} \tag{S.94}$$

$$= \begin{pmatrix} \log 2 \\ \log 4 \end{pmatrix} + \begin{pmatrix} \log 32 \\ \log 48 \end{pmatrix} = \begin{pmatrix} \log 64 \\ \log 192 \end{pmatrix} \tag{S.95}$$

The value x_1 for which $\gamma^*(x_1)$ is largest is thus $\hat{x}_1 = 1$. Plugging $\hat{x}_1 = 1$ into the back-tracking function $\lambda^*_{\phi_C \to x_1}(x_1)$ gives

$$(\hat{x}_1, \hat{x}_2, \hat{x}_3) = \operatorname*{argmax}_{x_1, x_2, x_3} p(x_1, x_2, x_3 | x_4 = 0, x_5 = 0) = (1, 1, 0). \tag{S.96}$$

In this low-dimensional example, we can verify the solution by computing the unnormalised pmf for all combinations of x_1, x_2, x_3. This is done in the following table where we start with the table for ϕ_C and then multiply-in the further factors ϕ_A, $\tilde{\phi}$ and ϕ_B.

x_1	x_2	x_3	ϕ_C	$\phi_C\phi_A$	$\phi_C\phi_A\tilde{\phi}$	$\phi_C\phi_A\tilde{\phi}\phi_B$
0	0	0	4	8	16	$16 \cdot 4$
1	0	0	2	8	16	$16 \cdot 4$
0	1	0	2	4	8	$8 \cdot 4$
1	1	0	6	24	48	$48 \cdot 4$
0	0	1	2	4	4	$4 \cdot 4$
1	0	1	6	24	24	$24 \cdot 4$
0	1	1	6	12	12	$12 \cdot 4$
1	1	1	4	16	16	$16 \cdot 4$

For example, for the column $\phi_c\phi_A$, we multiply each value of $\phi_C(x_1, x_2, x_3)$ by $\phi_A(x_1)$, so that the rows with $x_1 = 0$ get multiplied by 2, and the rows with $x_1 = 1$ by 4.

The maximal value in the final column is achieved for $x_1 = 1, x_2 = 1, x_3 = 0$, in line with the result above (and $48 \cdot 4 = 192$). Since $\phi_B(x_2)$ is a constant, being equal to 4 for all values of x_2, we could have ignored it in the computation. The formal reason for this is that since the model is unnormalised, we are allowed to rescale each factor by an arbitrary (factor-dependent) *constant*. This operation does not change the model. So we could divide ϕ_B by 4 which would give a value of 1, so that the factor can indeed be ignored.

(c) *Compute* $\operatorname{argmax}_{x_1,\ldots,x_5} p(x_1, \ldots, x_5)$ *via max-sum message passing with* x_1 *as root.*

Solution. As discussed in the solution to the answer above, we can drop factor $\phi_B(x_2)$ since it takes the same value for all x_2. Moreover, we can rescale the individual factors by a constant so they are more amenable to calculations by hand. We normalise them such that the largest value is one, which gives the following factors. Note that this is entirely optional.

x_1	ϕ_A
0	1
1	2

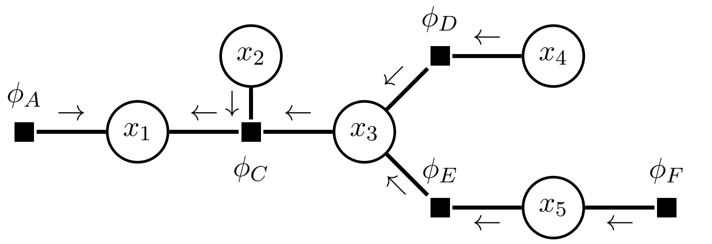
x_1	x_2	x_3	ϕ_C
0	0	0	2
1	0	0	1
0	1	0	1
1	1	0	3
0	0	1	1
1	0	1	3
0	1	1	3
1	1	1	2

x_3	x_4	ϕ_D
0	0	4
1	0	1
0	1	1
1	1	3

x_3	x_5	ϕ_E
0	0	1
1	0	2
0	1	2
1	1	1

x_5	ϕ_F
0	1
1	8

The factor graph without ϕ_B together with the messages that we need to compute is:



The leaf (log) messages are (using vector notation where the top element corresponds to $x_i = 0$ and the bottom one to $x_i = 1$):

$$\boldsymbol{\lambda}_{\phi_A \to x_1} = \begin{pmatrix} 0 \\ \log 2 \end{pmatrix} \quad \boldsymbol{\lambda}_{x_2 \to \phi_C} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \boldsymbol{\lambda}_{x_4 \to \phi_D} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \boldsymbol{\lambda}_{\phi_F \to x_5} = \begin{pmatrix} 0 \\ \log 8 \end{pmatrix} \tag{S.97}$$

The variable node x_5 only has one incoming edge so that $\boldsymbol{\lambda}_{x_5 \to \phi_E} = \boldsymbol{\lambda}_{\phi_F \to x_5}$. The message $\lambda_{\phi_E \to x_3}(x_3)$ equals

$$\lambda_{\phi_E \to x_3}(x_3) = \max_{x_5} \log \phi_E(x_3, x_5) + \lambda_{x_5 \to \phi_E}(x_5) \tag{S.98}$$

Writing out $\log \phi_E(x_3, x_5) + \lambda_{x_5 \to \phi_E}(x_5)$ for all x_5 as a function of x_3 we have

x_5	$\log \phi_E(x_3 = 0, x_5) + \lambda_{x_5 \to \phi_E}(x_5)$
0	$\log 1 + 0 = 0$
1	$\log 2 + \log 8 = \log 16$

x_5	$\log \phi_E(x_3 = 1, x_5) + \lambda_{x_5 \to \phi_E}(x_5)$
0	$\log 2 + 0 = \log 2$
1	$\log 1 + \log 8 = \log 8$

Taking the maximum over x_5 as a function of x_3, we obtain

$$\boldsymbol{\lambda}_{\phi_E \to x_3} = \begin{pmatrix} \log 16 \\ \log 8 \end{pmatrix} \tag{S.99}$$

and the backtracking function that indicates the maximiser $\hat{x}_5 = \text{argmax}_{x_5} \log \phi_E(x_3, x_5) + \lambda_{x_5 \to \phi_E}(x_5)$ as a function of x_3 equals

$$\lambda^*_{\phi_E \to x_3}(x_3) = \begin{cases} \hat{x}_5 = 1 & \text{if } x_3 = 0 \\ \hat{x}_5 = 1 & \text{if } x_3 = 1 \end{cases} \tag{S.100}$$

We perform the same kind of operation for $\lambda_{\phi_D \to x_3}(x_3)$

$$\lambda_{\phi_D \to x_3}(x_3) = \max_{x_4} \log \phi_D(x_3, x_4) + \lambda_{x_4 \to \phi_D}(x_4) \tag{S.101}$$

Since $\lambda_{x_4 \to \phi_D}(x_4) = 0$ for all x_4, the table with all values of $\log \phi_D(x_3, x_4) + \lambda_{x_4 \to \phi_D}(x_4)$ is

x_3	x_4	$\log \phi_D(x_3, x_4) + \lambda_{x_4 \to \phi_D}(x_4)$
0	0	$\log 4 + 0 = \log 4$
1	0	$\log 1 + 0 = 0$
0	1	$\log 1 + 0 = 0$
1	1	$\log 3 + 0 = \log 3$

Taking the maximum over x_4 as a function of x_3 we thus obtain

$$\boldsymbol{\lambda}_{\phi_D \to x_3} = \begin{pmatrix} \log 4 \\ \log 3 \end{pmatrix} \tag{S.102}$$

and the backtracking function that indicates the maximiser $\hat{x}_4 = \text{argmax}_{x_4} \log \phi_D(x_3, x_4) + \lambda_{x_4 \to \phi_D}(x_4)$ as a function of x_3 equals

$$\lambda^*_{\phi_D \to x_3}(x_3) = \begin{cases} \hat{x}_4 = 0 & \text{if } x_3 = 0 \\ \hat{x}_4 = 1 & \text{if } x_3 = 1 \end{cases} \tag{S.103}$$

For the message $\lambda_{x_3 \to \phi_C}(x_3)$ we add together the messages $\lambda_{\phi_E \to x_3}(x_3)$ and $\lambda_{\phi_D \to x_3}(x_3)$ which gives

$$\boldsymbol{\lambda}_{x_3 \to \phi_C} = \begin{pmatrix} \log 16 + \log 4 \\ \log 8 + \log 3 \end{pmatrix} = \begin{pmatrix} \log 64 \\ \log 24 \end{pmatrix} \tag{S.104}$$

Next we compute the message $\lambda_{\phi_C \to x_1}(x_1)$ by maximising over x_2 and x_3,

$$\lambda_{\phi_C \to x_1}(x_1) = \max_{x_2, x_3} \log \phi_C(x_1, x_2, x_3) + \lambda_{x_2 \to \phi_C}(x_2) + \lambda_{x_3 \to \phi_C}(x_3) \tag{S.105}$$

Since $\lambda_{x_2 \to \phi_C}(x_2) = 0$, the problem becomes

$$\lambda_{\phi_C \to x_1}(x_1) = \max_{x_2, x_3} \log \phi_C(x_1, x_2, x_3) + \lambda_{x_3 \to \phi_C}(x_3) \tag{S.106}$$

Building on the table for ϕ_C, we form a table with all values of $\log \phi_C(x_1, x_2, x_3) + \lambda_{x_3 \to \phi_C}(x_3)$

x_1	x_2	x_3	$\log \phi_C(x_1, x_2, x_3) + \lambda_{x_3 \to \phi_C}(x_3)$
0	0	0	$\log 2 + \log 64 = \mathbf{\log 128}$
1	0	0	$0 + \log 64 = \log 64$
0	1	0	$0 + \log 64 = \log 64$
1	1	0	$\log 3 + \log 64 = \mathbf{\log 192}$
0	0	1	$\log 24$
1	0	1	$\log 3 + \log 24 = \log 72$
0	1	1	$\log 3 + \log 24 = \log 72$
1	1	1	$\log 2 + \log 24 = \log 48$

The maximal value as a function of x_1 are highlighted in the table, which gives the message

$$\boldsymbol{\lambda}_{\phi_C \to x_1} = \begin{pmatrix} \log 128 \\ \log 192 \end{pmatrix} \tag{S.107}$$

and the backtracking function

$$\lambda^*_{\phi_C \to x_1}(x_1) = \begin{cases} (\hat{x}_2 = 0, \hat{x}_3 = 0) & \text{if } x_1 = 0 \\ (\hat{x}_2 = 1, \hat{x}_3 = 0) & \text{if } x_1 = 1 \end{cases} \tag{S.108}$$

We now have all incoming messages to the assigned root node x_1. *Ignoring the normalising constant*, we obtain

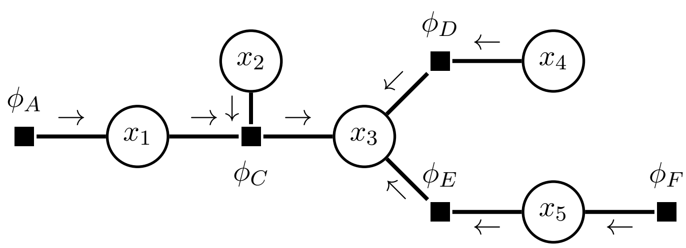
$$\boldsymbol{\gamma} = \begin{pmatrix} \gamma^*(x_1 = 0) \\ \gamma^*(x_1 = 1) \end{pmatrix} = \begin{pmatrix} 0 + \log 128 \\ \log 2 + \log 192 \end{pmatrix} \tag{S.109}$$

We can now start the backtracking to compute the desired $\operatorname{argmax}_{x_1,\ldots,x_5} p(x_1, \ldots, x_5)$. Starting at the root we have $\hat{x}_1 = \operatorname{argmax}_{x_1} \gamma^*(x_1) = 1$. Plugging this value into the look-up table $\lambda^*_{\phi_C \to x_1}(x_1)$, we obtain $(\hat{x}_2 = 1, \hat{x}_3 = 0)$. With the look-up table $\lambda^*_{\phi_E \to x_3}(x_3)$ we find $\hat{x}_5 = 1$ and $\lambda^*_{\phi_D \to x_3}(x_3)$ gives $\hat{x}_4 = 0$ so that overall

$$\operatorname*{argmax}_{x_1,\ldots,x_5} p(x_1, \ldots, x_5) = (1, 1, 0, 0, 1). \tag{S.110}$$

(d) *Compute* $\operatorname{argmax}_{x_1,\ldots,x_5} p(x_1, \ldots, x_5)$ *via max-sum message passing with* x_3 *as root.*

Solution. With x_3 as root, we need the following messages:



The following messages are the same as when x_1 was the root:

$$\boldsymbol{\lambda}_{\phi_D \to x_3} = \begin{pmatrix} \log 4 \\ \log 3 \end{pmatrix} \quad \boldsymbol{\lambda}_{\phi_E \to x_3} = \begin{pmatrix} \log 16 \\ \log 8 \end{pmatrix} \quad \boldsymbol{\lambda}_{\phi_A \to x_1} = \begin{pmatrix} 0 \\ \log 2 \end{pmatrix} \quad \boldsymbol{\lambda}_{x_2 \to \phi_C} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{S.111}$$

Since x_1 has only one incoming message, we further have

$$\boldsymbol{\lambda}_{x_1 \to \phi_C} = \boldsymbol{\lambda}_{\phi_A \to x_1} = \begin{pmatrix} 0 \\ \log 2 \end{pmatrix}. \tag{S.112}$$

We next compute $\lambda_{\phi_C \to x_3}(x_3)$,

$$\lambda_{\phi_C \to x_3}(x_3) = \max_{x_1, x_2} \log \phi_C(x_1, x_2, x_3) + \lambda_{x_1 \to \phi_C}(x_1) + \lambda_{x_2 \to \phi_C}(x_2). \tag{S.113}$$

We first form a table for $\log \phi_C(x_1, x_2, x_3) + \lambda_{x_1 \to \phi_C}(x_1) + \lambda_{x_2 \to \phi_C}(x_2)$ noting that $\lambda_{x_2 \to \phi_C}(x_2) = 0$

x_1	x_2	x_3	$\log \phi_C(x_1, x_2, x_3) + \lambda_{x_1 \to \phi_C}(x_1) + \lambda_{x_2 \to \phi_C}(x_2)$
0	0	0	$\log 2 + 0 = \log 2$
1	0	0	$0 + \log 2 = \log 2$
0	1	0	$0 + 0 = 0$
1	1	0	$\log 3 + \log 2 = \mathbf{log\ 6}$
0	0	1	$0 + 0 = 0$
1	0	1	$\log 3 + \log 2 = \mathbf{log\ 6}$
0	1	1	$\log 3 + 0 = \log 3$
1	1	1	$\log 2 + \log 2 = \log 4$

The maximal value as a function of x_3 are highlighted in the table, which gives the message

$$\boldsymbol{\lambda}_{\phi_C \to x_3} = \begin{pmatrix} \log 6 \\ \log 6 \end{pmatrix} \tag{S.114}$$

and the backtracking function

$$\lambda^*_{\phi_C \to x_3}(x_3) = \begin{cases} (\hat{x}_1 = 1, \hat{x}_2 = 1) & \text{if } x_3 = 0 \\ (\hat{x}_1 = 1, \hat{x}_2 = 0) & \text{if } x_3 = 1 \end{cases} \tag{S.115}$$

We have now all incoming messages for x_3 and can compute $\gamma^*(x_3)$ up the normalising constant $-\log Z$ (which is not needed if we are interested in the argmax only:

$$\boldsymbol{\gamma} = \begin{pmatrix} \gamma^*(x_3 = 0) \\ \gamma^*(x_3 = 1) \end{pmatrix} = \boldsymbol{\lambda}_{\phi_C \to x_3} + \boldsymbol{\lambda}_{\phi_D \to x_3} + \boldsymbol{\lambda}_{\phi_E \to x_3} \tag{S.116}$$

$$= \begin{pmatrix} \log 6 + \log 4 + \log 16 = \log 384 \\ \log 6 + \log 3 + \log 8 = \log 144 \end{pmatrix} \tag{S.117}$$

We can now start the backtracking which gives: $\hat{x}_3 = 0$, so that $\lambda^*_{\phi_C \to x_3}(0) = (\hat{x}_1 = 1, \hat{x}_2 = 1)$. The backtracking functions $\lambda^*_{\phi_E \to x_3}(x_3)$ and $\lambda^*_{\phi_D \to x_3}(x_3)$ are the same for question (c), which gives $\lambda^*_{\phi_E \to x_3}(0) = \hat{x}_5 = 1$ and $\lambda^*_{\phi_D \to x_3}(0) = \hat{x}_4 = 0$. Hence, overall, we find

$$\underset{x_1, \ldots, x_5}{\operatorname{argmax}}\, p(x_1, \ldots, x_5) = (1, 1, 0, 0, 1). \tag{S.118}$$

Note that this matches the result from question (c) where x_1 was the root. This is because the output of the max-sum algorithm is invariant to the choice of the root.

Exercise 5. ***Inference for two linearly dependent Gaussian RVs***

You have prior knowledge that an unknown variable $X \sim N(0, 1)$. You can make an observation of the variable Y which is related to X by $Y = wX + \mu_y + \epsilon$, with w and μ constants, and $\epsilon \sim N(0, \sigma^2)$ independent of X. The graphical model is $X \to Y$. (The factor w might arise e.g. because you want to measure X in centimeters, but your ruler is in inches. μ_y may arise due to an offset between the origins of the coordinates in the X and Y spaces.)

(a) *Show that $\mathbb{E}[Y] = \mu_y$.*

Solution.

$$\mathbb{E}[Y] = w\mathbb{E}[X] + \mathbb{E}[\mu_y] + \mathbb{E}[\epsilon] = 0 + \mu_y + 0 = \mu_y. \tag{S.119}$$

(b) *You now want to make inferences for X given the observation $Y = y$. The conditional distribution $p(X = x|Y = y)$ is Gaussian. Compute its posterior mean and variance.*
HINT: You cause the conditioning formula for Gaussians given in the slides, or manipulate the expressions for the Gaussians $p(x)$ and $p(y|x)$ directly.

Solution. One way to solve this problem is to use the conditioning formula for Gaussians given in the slides. As $X \sim N(0, 1)$, we have $\mathbb{E}[X] = 0$, and from above $\mathbb{E}[Y] = \mu_y$. We now need to compute the elements of the 2×2 covariance matrix Σ. We have that $\Sigma_{xx} = \mathbb{V}(X) = 1$.

To complete the covariance matrix, we need the entries $\text{covar}(X, Y)$ and $\mathbb{V}(Y)$.

$$\begin{aligned}
\Sigma_{xy} &= \text{covar}(X, Y) = \mathbb{E}[(X - 0)(Y - \mu_y)] = \mathbb{E}[X(wX + \epsilon)] = w\mathbb{V}(X) + 0 = w. \\
\Sigma_{yy} &= \mathbb{V}(Y) = \mathbb{E}[(Y - \mu_y)^2] = \mathbb{E}[(wX + \epsilon)^2] = w^2\mathbb{V}(X) + \mathbb{V}(\epsilon) = w^2 + \sigma^2.
\end{aligned}$$

The conditioning formula given in the slides implies:

$$\begin{aligned}
\mu^c_{x|y} &= \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y), \\
\Sigma^c_{x|y} &= \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx},
\end{aligned}$$

where $\Sigma^c_{x|y}$ denotes the conditional variance of x given y. We rename $(\mu^c_{x|y}, \Sigma^c_{x|y})$ as (μ_c, σ^2_c) for simplicity below. Substituting in and simplifying, we obtain

$$\mu_c = 0 + \frac{w}{w^2 + \sigma^2}(y - \mu_y) = \frac{w}{w^2 + \sigma^2}(y - \mu_y) \tag{S.120}$$

$$\sigma^2_c = 1 - \frac{w^2}{w^2 + \sigma^2} = \frac{w^2 + \sigma^2}{w^2 + \sigma^2} - \frac{w^2}{(w^2 + \sigma^2)} = \frac{\sigma^2}{w^2 + \sigma^2}. \tag{S.121}$$

If there was no noise we would write $x = (y - \mu_y)/w$, and this is indeed what would be obtained for μ_c in the limit $\sigma^2 \to 0$ in eq. S.120. Notice that the posterior variance σ^2_c is less than the prior variance, 1, so the observation has reduced uncertainty about X.

An alternative approach is to manipulate the expressions for the Gaussians $p(x)$ and $p(y|x)$ directly. We have $p(x|y) \propto p(x, y)$ when y is fixed. Hence

$$\begin{aligned}
p(x|y) &\propto p(x, y) = p(x)p(y|x) && \text{(S.122)} \\
&\propto \exp -\frac{1}{2}x^2 \cdot \exp -\frac{1}{2}\frac{(y - (wx + \mu_y))^2}{\sigma^2}. && \text{(S.123)}
\end{aligned}$$

Note that the normalization factors can be omitted. The above equation is (for fixed y) a Gaussian distribution in x. We first extract the quadratic form from within the exponents (omitting the 1/2 factor) to give

$$Q(x) = x^2 + \frac{(wx + \mu_y - y)^2}{\sigma^2} \tag{S.124}$$
$$= x^2(1 + \frac{w^2}{\sigma^2}) + 2x\frac{w(\mu_y - y)}{\sigma^2} + const \tag{S.125}$$

where *const* includes any terms that do not involve powers of x.
The conditional Gaussian in x has mean μ_c and variance σ_c^2. Its quadratic form is

$$Q_c(x) = \frac{(x - \mu_c)^2}{\sigma_c^2}. \tag{S.126}$$

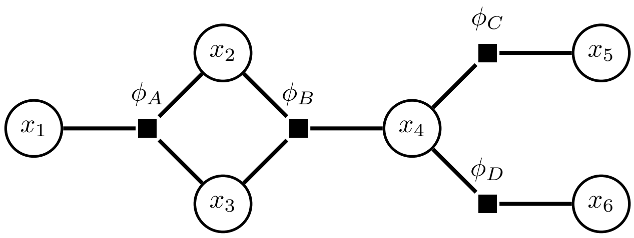
By comparing the coefficients of x^2 and x we obtain

$$\frac{1}{\sigma_c^2} = \frac{\sigma^2 + w^2}{\sigma^2} \Rightarrow \sigma_c^2 = \frac{\sigma^2}{\sigma^2 + w^2}, \tag{S.127}$$
$$\frac{\mu_c}{\sigma_c^2} = \frac{w}{\sigma^2}(y - \mu_y) \Rightarrow \mu_c = \frac{w}{\sigma^2 + w^2}(y - \mu_y), \tag{S.128}$$

in agreement with eqs. S.120 and S.121.

Exercise 6. *Choice of elimination order in factor graphs*

Consider the following factor graph, which contains a loop:



Let all variables be binary, $x_i \in \{0, 1\}$, and the factors be defined as follows:

x_1	x_2	x_3	ϕ_A
0	0	0	4
1	0	0	2
0	1	0	2
1	1	0	6
0	0	1	2
1	0	1	6
0	1	1	6
1	1	1	4

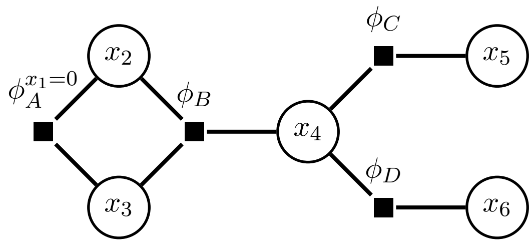
x_2	x_3	x_4	ϕ_B
0	0	0	2
1	0	0	2
0	1	0	4
1	1	0	2
0	0	1	6
1	0	1	8
0	1	1	4
1	1	1	2

x_4	x_5	ϕ_C
0	0	8
1	0	2
0	1	2
1	1	6

x_4	x_6	ϕ_D
0	0	3
1	0	6
0	1	6
1	1	3

(a) *Draw the factor graph corresponding to $p(x_2, x_3, x_4, x_5 \mid x_1 = 0, x_6 = 1)$ and give the tables defining the new factors $\phi_A^{x_1=0}(x_2, x_3)$ and $\phi_D^{x_6=1}(x_4)$ that you obtain.*

Solution. First condition on $x_1 = 0$:
Factor node $\phi_A(x_1, x_2, x_3)$ depends on x_1, thus we create a new factor $\phi_A^{x_1=0}(x_2, x_3)$ from the table for ϕ_A using the rows where $x_1 = 0$.

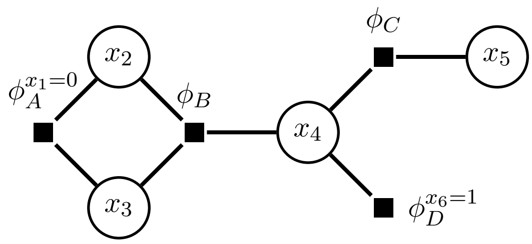


	x_1	x_2	x_3	ϕ_A
$\rightarrow$	0	0	0	4
	1	0	0	2
$\rightarrow$	0	1	0	2
	1	1	0	6
$\rightarrow$	0	0	1	2
	1	0	1	6
$\rightarrow$	0	1	1	6
	1	1	1	4

so that

x_2	x_3	$\phi_A^{x_1=0}$
0	0	4
1	0	2
0	1	2
1	1	6

Next condition on $x_6 = 1$:
Factor node $\phi_D(x_4, x_6)$ depends on x_6, thus we create a new factor $\phi_D^{x_6=1}(x_4)$ from the table for ϕ_D using the rows where $x_6 = 1$.



	x_4	x_6	ϕ_D
	0	0	3
	1	0	6
$\rightarrow$	0	1	6
$\rightarrow$	1	1	3

so that

x_4	$\phi_D^{x_6=1}$
0	6
1	3

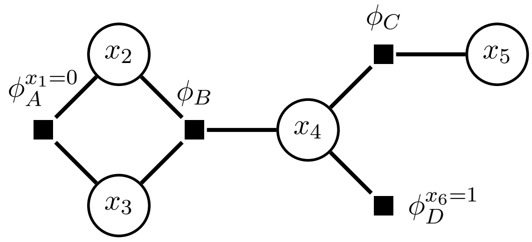
(b) *Find $p(x_2 \mid x_1 = 0, x_6 = 1)$ using the elimination ordering (x_4, x_5, x_3):*

(i) *Draw the graph for $p(x_2, x_3, x_5 \mid x_1 = 0, x_6 = 1)$ by marginalising x_4*
Compute the table for the new factor $\tilde{\phi}_4(x_2, x_3, x_5)$

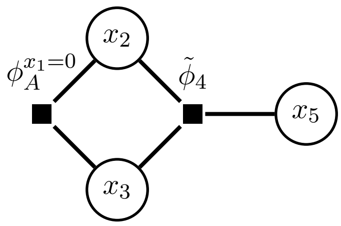
(ii) *Draw the graph for $p(x_2, x_3 \mid x_1 = 0, x_6 = 1)$ by marginalising x_5*
Compute the table for the new factor $\tilde{\phi}_{45}(x_2, x_3)$

(iii) *Draw the graph for $p(x_2 \mid x_1 = 0, x_6 = 1)$ by marginalising x_3*
Compute the table for the new factor $\tilde{\phi}_{453}(x_2)$

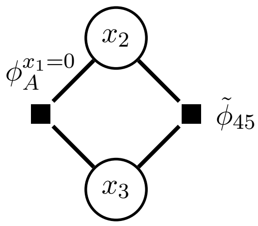
Solution. Starting with the factor graph for $p(x_2, x_3, x_4, x_5 \mid x_1 = 0, x_6 = 1)$



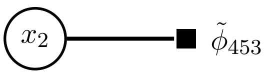
Marginalising x_4 combines the three factors ϕ_B, ϕ_C and $\phi_D^{x_6=1}$



Marginalising x_5 modifies the factor $\tilde{\phi}_4$



Marginalising x_3 combines the factors $\phi_A^{x_1=0}$ and $\tilde{\phi}_{45}$



We now compute the tables for the new factors $\tilde{\phi}_4$, $\tilde{\phi}_{45}$, $\tilde{\phi}_{453}$.
First find $\tilde{\phi}_4(x_2, x_3, x_5)$

x_2	x_3	x_4	ϕ_B
0	0	0	2
1	0	0	2
0	1	0	4
1	1	0	2
0	0	1	6
1	0	1	8
0	1	1	4
1	1	1	2

x_4	x_5	ϕ_C
0	0	8
1	0	2
0	1	2
1	1	6

x_4	$\phi_D^{x_6=1}$
0	6
1	3

so that $\phi_*(x_2, x_3, x_4, x_5) = \phi_B(x_2, x_3, x_4)\phi_C(x_4, x_5)\phi_D^{x_6=1}(x_4)$ equals

x_2	x_3	x_4	x_5	$\phi_*(x_2, x_3, x_4, x_5)$
0	0	0	0	2 * 8 * 6
1	0	0	0	2 * 8 * 6
0	1	0	0	4 * 8 * 6
1	1	0	0	2 * 8 * 6
0	0	1	0	6 * 2 * 3
1	0	1	0	8 * 2 * 3
0	1	1	0	4 * 2 * 3
1	1	1	0	2 * 2 * 3
0	0	0	1	2 * 2 * 6
1	0	0	1	2 * 2 * 6
0	1	0	1	4 * 2 * 6
1	1	0	1	2 * 2 * 6
0	0	1	1	6 * 6 * 3
1	0	1	1	8 * 6 * 3
0	1	1	1	4 * 6 * 3
1	1	1	1	2 * 6 * 3

and

x_2	x_3	x_5	$\sum_{x_4} \phi_B(x_2, x_3, x_4)\phi_C(x_4, x_5)\phi_D^{x_6=1}(x_4)$		$\tilde{\phi}_4$
0	0	0	(2 * 8 * 6) + (6 * 2 * 3)	=	132
1	0	0	(2 * 8 * 6) + (8 * 2 * 3)	=	144
0	1	0	(4 * 8 * 6) + (4 * 2 * 3)	=	216
1	1	0	(2 * 8 * 6) + (2 * 2 * 3)	=	108
0	0	1	(2 * 2 * 6) + (6 * 6 * 3)	=	132
1	0	1	(2 * 2 * 6) + (8 * 6 * 3)	=	168
0	1	1	(4 * 2 * 6) + (4 * 6 * 3)	=	120
1	1	1	(2 * 2 * 6) + (2 * 6 * 3)	=	60

Next find $\tilde{\phi}_{45}(x_2, x_3)$

x_2	x_3	x_5	$\tilde{\phi}_4$
0	0	0	132
1	0	0	144
0	1	0	216
1	1	0	108
0	0	1	132
1	0	1	168
0	1	1	120
1	1	1	60

so that

x_2	x_3	$\sum_{x_5} \tilde{\phi}_4(x_2, x_3, x_5)$		$\tilde{\phi}_{45}$
0	0	132 + 132	=	264
1	0	144 + 168	=	312
0	1	216 + 120	=	336
1	1	108 + 60	=	168

Finally find $\tilde{\phi}_{453}(x_2)$

x_2	x_3	$\phi_A^{x_1=0}$
0	0	4
1	0	2
0	1	2
1	1	6

x_2	x_3	$\tilde{\phi}_{45}$
0	0	264
1	0	312
0	1	336
1	1	168

so that

x_2	$\sum_{x_3} \tilde{\phi}_{45}(x_2, x_3)\phi_A^{x_1=0}(x_2, x_3)$		$\tilde{\phi}_{453}$
0	(4 * 264) + (2 * 336)	=	1728
1	(2 * 312) + (6 * 168)	=	1632

The normalising constant is $Z = 1728 + 1632$. Our conditional marginal is thus:

$$p(x_2 \mid x_1 = 0, x_6 = 1) = \begin{pmatrix} 1728/Z \\ 1632/Z \end{pmatrix} = \begin{pmatrix} 0.514 \\ 0.486 \end{pmatrix} \tag{S.129}$$

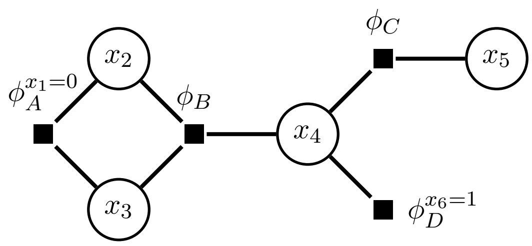
(c) *Now determine $p(x_2 \mid x_1 = 0, x_6 = 1)$ with the elimination ordering (x_5, x_4, x_3):*

(i) *Draw the graph for $p(x_2, x_3, x_4, \mid x_1 = 0, x_6 = 1)$ by marginalising x_5*
Compute the table for the new factor $\tilde{\phi}_5(x_4)$

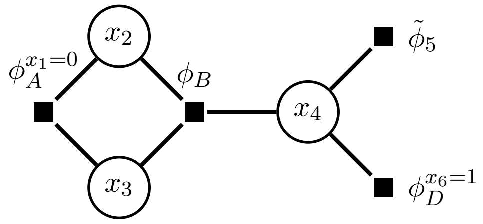
(ii) *Draw the graph for $p(x_2, x_3 \mid x_1 = 0, x_6 = 1)$ by marginalising x_4*
Compute the table for the new factor $\tilde{\phi}_{54}(x_2, x_3)$

(iii) *Draw the graph for $p(x_2 \mid x_1 = 0, x_6 = 1)$ by marginalising x_3*
Compute the table for the new factor $\tilde{\phi}_{543}(x_2)$

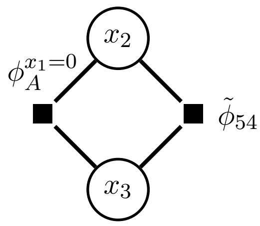
Solution. Starting with the factor graph for $p(x_2, x_3, x_4, x_5 \mid x_1 = 0, x_6 = 1)$



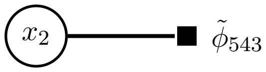
Marginalising x_5 modifies the factor ϕ_C



Marginalising x_4 combines the three factors ϕ_B, $\tilde{\phi}_5$ and $\phi_D^{x_6=1}$



Marginalising x_3 combines the factors $\phi_A^{x_1=0}$ and $\tilde{\phi}_{54}$



We now compute the tables for the new factors $\tilde{\phi}_5$, $\tilde{\phi}_{54}$, and $\tilde{\phi}_{543}$.
First find $\tilde{\phi}_5(x_4)$

x_4	x_5	ϕ_C
0	0	8
1	0	2
0	1	2
1	1	6

so that

x_4	$\sum_{x_5} \phi_C(x_4, x_5)$		$\tilde{\phi}_5$
0	$8 + 2$	$=$	10
1	$2 + 6$	$=$	8

Next find $\tilde{\phi}_{54}(x_2, x_3)$

x_2	x_3	x_4	ϕ_B
0	0	0	2
1	0	0	2
0	1	0	4
1	1	0	2
0	0	1	6
1	0	1	8
0	1	1	4
1	1	1	2

x_4	$\tilde{\phi}_5$
0	10
1	8

x_4	$\phi_D^{x_6=1}$
0	6
1	3

so that $\phi_*(x_2, x_3, x_4) = \phi_B(x_2, x_3, x_4)\tilde{\phi}_5(x_4)\phi_D^{x_6=1}(x_4)$ equals

x_2	x_3	x_4	$\phi_*(x_2, x_3, x_4)$
0	0	0	$2 * 10 * 6$
1	0	0	$2 * 10 * 6$
0	1	0	$4 * 10 * 6$
1	1	0	$2 * 10 * 6$
0	0	1	$6 * 8 * 3$
1	0	1	$8 * 8 * 3$
0	1	1	$4 * 8 * 3$
1	1	1	$2 * 8 * 3$

and

x_2	x_3	$\sum_{x_4} \phi_B(x_2, x_3, x_4)\tilde{\phi}_5(x_4)\phi_D^{x_6=1}(x_4)$		$\tilde{\phi}_{54}$
0	0	$(2 * 10 * 6) + (6 * 8 * 3)$	$=$	264
1	0	$(2 * 10 * 6) + (8 * 8 * 3)$	$=$	312
0	1	$(4 * 10 * 6) + (4 * 8 * 3)$	$=$	336
1	1	$(2 * 10 * 6) + (2 * 8 * 3)$	$=$	168

Finally find $\tilde{\phi}_{543}(x_2)$

x_2	x_3	$\phi_A^{x_1=0}$
0	0	4
1	0	2
0	1	2
1	1	6

x_2	x_3	$\tilde{\phi}_{54}$
0	0	264
1	0	312
0	1	336
1	1	168

so that

x_2	$\sum_{x_3} \tilde{\phi}_{54}(x_2, x_3)\phi_A^{x_1=0}(x_2, x_3)$		$\tilde{\phi}_{543}$
0	(4 * 264) + (2 * 336)	=	1728
1	(2 * 312) + (6 * 168)	=	1632

As with the ordering in the previous part, we should come to the same result for our conditional marginal distribution.The normalising constant is $Z = 1728 + 1632$, so that the conditional marginal is

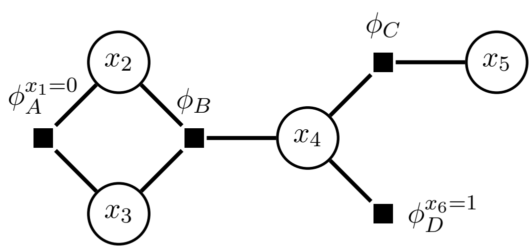
$$p(x_2 \mid x_1 = 0, x_6 = 1) = \begin{pmatrix} 1728/Z \\ 1632/Z \end{pmatrix} = \begin{pmatrix} 0.514 \\ 0.486 \end{pmatrix} \tag{S.130}$$

(d) *Which variable ordering, (x_4, x_5, x_3) or (x_5, x_4, x_3) do you prefer?*

Solution. The ordering (x_5, x_4, x_3) is cheaper and should be preferred over the ordering (x_4, x_5, x_3) .

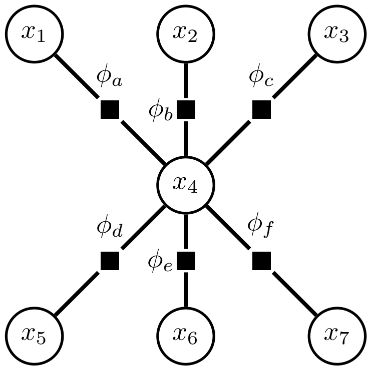
The reason for the difference in the cost is that x_4 has three neighbours in the factor graph for $p(x_2, x_3, x_4, x_5 \mid x_1 = 0, x_6 = 1)$. However, after elimination of x_5, which has only one neighbour, x_4 has only two neighbours left. Eliminating variables with more neighbours leads to larger (temporary) factors and hence a larger cost. We can see this from the tables that were generated during the computation (or numbers that we needed to add together): for the ordering (x_4, x_5, x_3), the largest table had 2^4 entries while for (x_5, x_4, x_3), it had 2^3 entries.

Choosing a reasonable variable ordering has a direct effect on the computational complexity of variable elimination. This effect becomes even more pronounced when the domain of our discrete variables has a size greater than 2 (binary variables), or if the variables are continuous.



Exercise 7. ***Choice of elimination order in factor graphs***

We would like to compute the marginal $p(x_1)$ by variable elimination for a joint pmf represented by the following factor graph. All variables x_i can take K different values.



(a) *A friend proposes the elimination order $x_4, x_5, x_6, x_7, x_3, x_2$, i.e. to do x_4 first and x_2 last. Explain why this is computationally inefficient.*

Solution. According to the factor graph, $p(x_1, \ldots, x_7)$ factorises as

$$p(x_1, \ldots, x_7) \propto \phi_a(x_1, x_4)\phi_b(x_2, x_4)\phi_c(x_3, x_4)\phi_d(x_5, x_4)\phi_e(x_6, x_4)\phi_f(x_7, x_4) \quad \text{(S.131)}$$

If we choose to eliminate x_4 first, i.e. compute

$$p(x_1, x_2, x_3, x_5, x_6, x_7) = \sum_{x_4} p(x_1, \ldots, x_7) \quad \text{(S.132)}$$

$$\propto \sum_{x_4} \phi_a(x_1, x_4)\phi_b(x_2, x_4)\phi_c(x_3, x_4)\phi_d(x_5, x_4)\phi_e(x_6, x_4)\phi_f(x_7, x_4) \quad \text{(S.133)}$$

we cannot pull any of the factors out of the sum since each of them depends on x_4. This means the cost to sum out x_4 for all combinations of the six variables $(x_1, x_2, x_3, x_5, x_6, x_7)$ is K^7. Moreover, the new factor

$$\tilde{\phi}(x_1, x_2, x_3, x_5, x_6, x_7) = \sum_{x_4} \phi_a(x_1, x_4)\phi_b(x_2, x_4)\phi_c(x_3, x_4)\phi_d(x_5, x_4)\phi_e(x_6, x_4)\phi_f(x_7, x_4) \quad \text{(S.134)}$$

does not factorise anymore so that subsequent variable eliminations will be expensive too.

(b) *Propose an elimination ordering that achieves $O(K^2)$ computational cost per variable elimination and explain why it does so.*

Solution. Any ordering where x_4 is eliminated last will do. At any stage, elimination of one of the variables x_2, x_3, x_5, x_6, x_7 is then a $O(K^2)$ operation. This is because e.g.

$$p(x_1, \ldots, x_6) = \sum_{x_7} p(x_1, \ldots, x_7) \quad \text{(S.135)}$$

$$\propto \phi_a(x_1, x_4)\phi_b(x_2, x_4)\phi_c(x_3, x_4)\phi_d(x_5, x_4)\phi_e(x_6, x_4) \underbrace{\sum_{x_7} \phi_f(x_7, x_4)}_{\tilde{\phi}_7(x_4)} \quad \text{(S.136)}$$

$$\propto \phi_a(x_1, x_4)\phi_b(x_2, x_4)\phi_c(x_3, x_4)\phi_d(x_5, x_4)\phi_e(x_6, x_4)\tilde{\phi}_7(x_4) \quad \text{(S.137)}$$

where computing $\tilde{\phi}_7(x_4)$ for all values of x_4 is $O(K^2)$. Further,

$$p(x_1, \ldots, x_5) = \sum_{x_6} p(x_1, \ldots, x_6) \tag{S.138}$$

$$\propto \phi_a(x_1, x_4)\phi_b(x_2, x_4)\phi_c(x_3, x_4)\phi_d(x_5, x_4)\tilde{\phi}_7(x_4) \sum_{x_6} \phi_e(x_6, x_4) \tag{S.139}$$

$$\propto \phi_a(x_1, x_4)\phi_b(x_2, x_4)\phi_c(x_3, x_4)\phi_d(x_5, x_4)\tilde{\phi}_7(x_4)\tilde{\phi}_6(x_4), \tag{S.140}$$

where computation of $\tilde{\phi}_6(x_4)$ for all values of x_4 is again $O(K^2)$. Continuing in this manner, one obtains

$$p(x_1, x_4) \propto \phi_a(x_1, x_4)\tilde{\phi}_2(x_4)\tilde{\phi}_3(x_4)\tilde{\phi}_5(x_4)\tilde{\phi}_6(x_4)\tilde{\phi}_7(x_4). \tag{S.141}$$

where each derived factor $\tilde{\phi}$ has $O(K^2)$ cost. Summing out x_4 and normalising the pmf is again a $O(K^2)$ operation.

