Exercise 6: 9 April 2019
CS-526 Learning Theory

Problem 1

1) The joint distribution is (up to normalisation factors of Gaussians)

$$p(y_1, \dots, y_m, x_1, \dots, x_m, w_1, \dots, w_p) \propto \prod_{i=1}^{m} e^{-\frac{1}{2\sigma^2}(y_i - \sum_{a=1}^{p} w_a x_i^a)^2} \prod_{i=1}^{m} P_0(x_i) \prod_{a=1}^{p} e^{-\alpha w_a^2}$$

2) Here the x_i is a parent of y_i (for all $i = 1, \dots, m$) and $w_1, \dots, w_p$ are parents of each $y_i, i = 1, \dots, m$.

3) The ML principle says that you maximize the log-likelihood $\log P(data \mid w_1, \dots, w_p)$. Since

$$P(data \mid w_1, \dots, w_p) \propto \prod_{i=1}^{m} e^{-\frac{1}{2\sigma^2}(y_i - \sum_{a=1}^{p} w_a x_i^a)^2} \prod_{i=1}^{m} P_0(x_i)$$

this is equivalent to minimising

$$\mathcal{E}_{data}(f) = \frac{1}{m} \sum_{i=1}^{m} (y_i - f(x_i))^2$$

over functions in the class $\mathcal{H} \ni f(x) = \sum_{a=1}^{p} w_a x^a$.

4) The posterior distribution is

$$\begin{aligned} P(w_1, \dots, w_p \mid data) &= \frac{p(y_1, \dots, y_m, x_1, \dots, x_m, w_1, \dots, w_p)}{\int \prod_{a=1}^{p} dw_a p(y_1, \dots, y_m, x_1, \dots, x_m, w_1, \dots, w_p)} \\ &= \frac{\prod_{i=1}^{m} e^{-\frac{1}{2\sigma^2}(y_i - \sum_{a=1}^{p} w_a x_i^a)^2} \prod_{a=1}^{p} e^{-\alpha w_a^2}}{\int \prod_{a=1}^{p} dw_a \prod_{i=1}^{m} e^{-\frac{1}{2\sigma^2}(y_i - \sum_{a=1}^{p} w_a x_i^a)^2} \prod_{a=1}^{p} e^{-\alpha w_a^2}} \end{aligned}$$

The MAP principle says you maximise the posterior which is equivalent to minimizing

$$\frac{1}{m} \sum_{i=1}^{m} (y_i - f(x_i))^2 + 2\alpha\sigma^2 \sum_{a=1}^{p} w_a^2$$

over the functions in the class $\mathcal{H} \ni f(x) = \sum_{a=1}^{p} w_a x^a$.

5) The optimal regression function is $f_{regr}(x) = \mathbb{E}_{w|data}\mathbb{E}_{y|x,w}[y]$. From the model it is clear that

$$\mathbb{E}_{y|x,w}[y] = \sum_{a=1}^{p} w_a x^a$$

Further average over the posterior gives

$$f_{regr}(x) = \sum_{a=1}^{p} \mathbb{E}_{w|data}[w_a] x^a$$

Problem 2

1) $a \perp\!\!\!\perp b|c$ because $p(a,b|c) = \frac{p(a,b,c)}{p(c)} = \frac{p(a|c)p(b|c)p(c)}{p(c)} = p(a|c)p(b|c)$. But a, b are not independent because $p(a,b) = \sum_c p(a|c)p(b|c) \neq p(a)p(b)$.

2) $a \perp\!\!\!\perp b|c$ because $p(a,b|c) = \frac{p(a,b,c)}{p(c)} = \frac{p(b|c)p(c|a)p(a)}{p(c)} = p(b|c)\frac{p(c|a)p(a)}{p(c)} = p(b|c)p(a|c)$. But a, b are not independent because $p(a,b) = \sum_c p(a)p(c|a)p(b|c) = p(a)p(b|a) \neq p(a)p(b)$.

3) $a \perp\!\!\!\perp b$ because

$$p(a,b) = \sum_{c,d} p(a,b,c,d) = \sum_{c,d} p(a)p(b)p(c|a,b)p(d|c) = p(a)p(b) \sum_{c,d} p(c|a,b)p(d|c) = p(a)p(b).$$

However, we don't have $a \perp\!\!\!\perp b|c$ because $p(a,b|c) = \frac{p(a)p(b)p(c|a,b))}{p(c)}$ cannot be decomposed.

Problem 3

The left hand side is

$$p(x_i|\mathbf{x}_{\sim i}) = \frac{p(\mathbf{x})}{\int dx_i \; p(\mathbf{x})} \tag{1}$$

where

$$p(\mathbf{x}) = p(x_i|\{x_v\}_{v\in \text{pa}(i)}) \prod_{k\in \text{child}(j)} p(x_j|\{x_v\}_{v\in \text{pa}(k)}) \prod_{\substack{l\neq i \\ l\neq \text{child}(i)}} p(x_l|\{x_v\}_{v\in \text{pa}(l)}).$$

The product $\prod_{\substack{l\neq i \\ l\neq \text{child}(i)}} p(x_l|\{x_v\}_{v\in \text{pa}(l)})$ is independent of x_i. It cancels with the same factor in the denominator of (1). So we have

$$p(x_i|\mathbf{x}_{\sim i}) = \frac{p(x_i|\{x_v\}_{v\in \text{pa}(i)}) \prod_{k\in \text{child}(j)} p(x_j|\{x_v\}_{v\in \text{pa}(k)})}{\int dx_i \; p(x_i|\{x_v\}_{v\in \text{pa}(i)}) \prod_{k\in \text{child}(j)} p(x_j|\{x_v\}_{v\in \text{pa}(k)})} \tag{2}$$

On the other hand, the right hand side is

$$p(x_i|\{x_v\}_{v\in \text{MB}(i)}) = \frac{p(x_i, \{x_v\}_{v\in \text{MB}(i)})}{\int dx_i \; p(x_i, \{x_v\}_{v\in \text{MB}(i)})} \tag{3}$$

where

$$
\begin{aligned}
&p(x_i, \{x_v\}_{v\in \mathrm{MB}(i)}) \\
&= \int d\mathbf{x}_{\sim i,\mathrm{MB}(i)}\ p(\mathbf{x}) \\
&= \int d\mathbf{x}_{\sim i,\mathrm{MB}(i)}\ p(x_i|\{x_v\}_{v\in \mathrm{pa}(i)}) \prod_{k\in \mathrm{child}(j)} p(x_j|\{x_v\}_{v\in \mathrm{pa}(k)}) \prod_{\substack{l\neq i\\ l\neq \mathrm{child}(i)}} p(x_l|\{x_v\}_{v\in \mathrm{pa}(l)}) \\
&= \ p(x_i|\{x_v\}_{v\in \mathrm{pa}(i)}) \prod_{k\in \mathrm{child}(j)} p(x_j|\{x_v\}_{v\in \mathrm{pa}(k)}) \Big[\int d\mathbf{x}_{\sim i,\mathrm{MB}(i)} \prod_{\substack{l\neq i\\ l\neq \mathrm{child}(i)}} p(x_l|\{x_v\}_{v\in \mathrm{pa}(l)}) \Big]
\end{aligned}
$$

We identify $\int d\mathbf{x}_{\sim i,\mathrm{MB}(i)} \prod_{\substack{l\neq i\\ l\neq \mathrm{child}(i)}} p(x_l|\{x_v\}_{v\in \mathrm{pa}(l)})$ independent of x_i. It cancels with the same factor in the denominator of (3). So (3) is reduced to the same expression as (2).

Problem 4 (Bishop, p.371 & 419, Exercise 8.7)

Using $\mathbb{E}[x_i] = \sum_{j\in \mathrm{pa}(i)} w_{ij}\mathbb{E}[x_j] + b_i$ gives

$$
\begin{aligned}
\mu_1 &= \sum_{j\in\emptyset} w_{1j}\mathbb{E}[x_j] + b_1 = b_1 \\
\mu_2 &= \sum_{j\in\{1\}} w_{2j}\mathbb{E}[x_j] = w_{21}b_1 + b_2 \\
\mu_3 &= \sum_{j\in\{2\}} w_{3j}\mathbb{E}[x_j] + b_3 = w_{32}(w_{21}b_1 + b_2) + b_3
\end{aligned}
$$

Using $\mathrm{cov}[x_i, x_j] = \sum_{k\in \mathrm{pa}(j)} w_{jk}\mathrm{cov}[x_i, x_k] + I_{ij}v_j$ for $i \leq j$ and $\mathrm{cov}[x_i, x_j] = \mathrm{cov}[x_j, x_i]$ gives

$$
\begin{aligned}
\mathrm{cov}[x_1, x_1] &= \sum_{k\in\emptyset} w_{1j}\mathrm{cov}[x_1, x_k] + v_1 = v_1 \\
\mathrm{cov}[x_1, x_2] &= \sum_{k\in\{1\}} w_{2j}\mathrm{cov}[x_1, x_k] = w_{21}v_1 \\
\mathrm{cov}[x_1, x_3] &= \sum_{k\in\{2\}} w_{3j}\mathrm{cov}[x_1, x_k] = w_{32}(w_{21}v_1) \\
\mathrm{cov}[x_2, x_2] &= \sum_{k\in\{1\}} w_{2j}\mathrm{cov}[x_2, x_k] + v_2 = w_{21}(w_{21}v_1) + v_2 \\
\mathrm{cov}[x_2, x_3] &= \sum_{k\in\{2\}} w_{3j}\mathrm{cov}[x_2, x_k] = w_{32}(w_{21}^2 v_1 + v_2) \\
\mathrm{cov}[x_3, x_3] &= \sum_{k\in\{2\}} w_{3j}\mathrm{cov}[x_3, x_k] + v_3 = w_{32}^2(w_{21}^2 v_1 + v_2) + v_3
\end{aligned}
$$

Problem 5 (Barber, p.75, Exercise 4.4)

1) First note that

$$p(\mathbf{h}|\mathbf{v}) \propto e^{(\mathbf{v}^\top \mathbf{W} + \mathbf{b}^\top)\mathbf{h}} = \prod_i e^{h_i(b_i + \sum_j W_{ji}v_j)}$$

So $p(\mathbf{h}|\mathbf{v}) = \prod_i p(h_i|\mathbf{v})$. Recall $h_i \in \{0,1\}$. Thus we have

$$p(h_i = 1|\mathbf{v}) = \frac{e^{b_i + \sum_j W_{ji}v_j}}{\sum_{h_i \in \{0,1\}} e^{h_i(b_i + \sum_j W_{ji}v_j)}} = \sigma\Big(b_i + \sum_j W_{ji}v_j\Big).$$

2)

$$p(\mathbf{v}|\mathbf{h}) = \prod_i p(v_i|\mathbf{h}), \qquad \text{with } p(v_i = 1|\mathbf{h}) = \sigma\Big(a_i + \sum_j W_{ij}h_j\Big)$$

3) No. Because the term $\mathbf{v}^\top \mathbf{W}\mathbf{h}$ in $p(\mathbf{v},\mathbf{h})$ introduces dependence between $\mathbf{v}$ and $\mathbf{h}$.

4) For a general $\mathbf{W}$ there is no known efficient way to compute Z efficiently. The dependence between $\mathbf{v}$ and $\mathbf{h}$ does not allow always decomposition of $p(\mathbf{v},\mathbf{h})$.

Problem 6 (Barber, p.77, Exercise 4.14)

We write

$$\begin{aligned}\phi_{ij}(x_i,x_j) &= e^{\ln \phi_{ij}(x_i,x_j)} \\ &= e^{\mathbb{I}(x_i=0,x_j=0)\ln\phi_{ij}(0,0) + \mathbb{I}(x_i=0,x_j=1)\ln\phi_{ij}(0,1) + \mathbb{I}(x_i=1,x_j=0)\ln\phi_{ij}(1,0) + \mathbb{I}(x_i=1,x_j=1)\ln\phi_{ij}(1,1)}\end{aligned}$$

With $x_i \in \{0,1\}$ we can replace $\mathbb{I}[\cdot]$ by

$$\begin{aligned}&\mathbb{I}(x_i=0,x_j=0) = (1-x_i)(1-x_j), &\qquad &\mathbb{I}(x_i=0,x_j=1) = (1-x_i)x_j, \\ &\mathbb{I}(x_i=1,x_j=0) = x_i(1-x_j), & &\mathbb{I}(x_i=1,x_j=1) = x_i x_j.\end{aligned}$$

So $\phi_{ij}(x_i,x_j)$ is in the form $e^{W_{ij}x_ix_j + b_ix_i + b_jx_j + \text{constant}}$ and $p(\mathbf{x}) = \frac{1}{Z'}e^{\sum_{ij\in\mathcal{E}} W_{ij}x_ix_j + \sum_i \deg(i) b_i x_i}$ is the Boltzmann machine.

Problem 7

Fix a subset $S \subseteq V$. We have:

$$p(\mathbf{x}_S, \mathbf{x}_{V\setminus S}) = p(\mathbf{x}) = \frac{1}{Z} \prod_{\substack{C'\in\mathcal{C}: \\ S\cap C' = \emptyset}} \psi_{C'}(\mathbf{x}_{C'}) \cdot \prod_{\substack{C\in\mathcal{C}: \\ S\cap C \neq \emptyset}} \psi_C(\mathbf{x}_C)\,;$$

$$p(\mathbf{x}_{V\setminus S}) = \sum_{\mathbf{x}_S} p(\mathbf{x}) = \frac{1}{Z} \prod_{\substack{C'\in\mathcal{C}: \\ S\cap C' = \emptyset}} \psi_{C'}(\mathbf{x}_{C'}) \cdot \left(\sum_{\mathbf{x}_S} \prod_{\substack{C\in\mathcal{C}: \\ S\cap C \neq \emptyset}} \psi_C(\mathbf{x}_C) \right).$$

Therefore, the conditional distribution of $\mathbf{x}_S$ given $\mathbf{x}_{V\setminus S}$ reads:

$$p(\mathbf{x}_S|\mathbf{x}_{V\setminus S}) = \frac{p(\mathbf{x}_S, \mathbf{x}_{V\setminus S})}{p(\mathbf{x}_{V\setminus S})} = \frac{\prod_{C:S\cap C\neq\emptyset} \psi_C(\mathbf{x}_C)}{\sum_{\widetilde{\mathbf{x}}_S} \prod_{C:S\cap C\neq\emptyset} \psi_C(\widetilde{\mathbf{x}}_C)}\,. \tag{4}$$

To write the denominator in the last equality, we implitcitly introduced $\widetilde{\mathbf{x}} = (\widetilde{\mathbf{x}}_S, \mathbf{x}_{V\setminus S})$, while $\mathbf{x} = (\mathbf{x}_S, \mathbf{x}_{V\setminus S})$.
Consider any maximal clique C such that $S \cap C \neq \emptyset$ and let $i \in S \cap C$. If $j \in C \setminus S$ then $j \in \partial S$ because $\{i, j\} \in E$ ($i \in C$ and C is a clique). Therefore $C \subseteq S \cup \partial S$. It follows:

$$p(\mathbf{x}_S, \mathbf{x}_{\partial S}) = \sum_{\mathbf{x}_{V\setminus(S\cup\partial S)}} p(\mathbf{x}) = \frac{1}{Z} \prod_{\substack{C\in\mathcal{C}:\\ S\cap C\neq\emptyset}} \psi_C(\mathbf{x}_C) \cdot \left(\sum_{\mathbf{x}_{V\setminus(S\cup\partial S)}} \prod_{\substack{C'\in\mathcal{C}:\\ S\cap C'=\emptyset}} \psi_{C'}(\mathbf{x}_{C'}) \right);$$

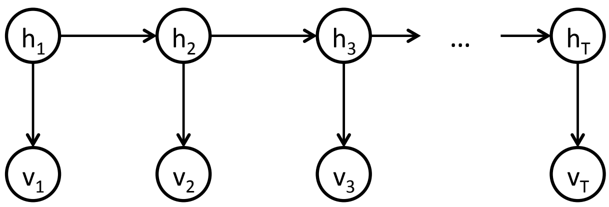
$$\begin{aligned} p(\mathbf{x}_{\partial S}) = \sum_{\mathbf{x}_{V\setminus(S\cup\partial S)}} p(\mathbf{x}_{V\setminus S}) &= \sum_{\mathbf{x}_{V\setminus(S\cup\partial S)}} \frac{1}{Z} \prod_{\substack{C'\in\mathcal{C}:\\ S\cap C'=\emptyset}} \psi_{C'}(\mathbf{x}_{C'}) \cdot \left(\sum_{\mathbf{x}_S} \prod_{\substack{C\in\mathcal{C}:\\ S\cap C\neq\emptyset}} \psi_C(\mathbf{x}_C) \right) \\ &= \frac{1}{Z} \left(\sum_{\mathbf{x}_{V\setminus(S\cup\partial S)}} \prod_{\substack{C'\in\mathcal{C}:\\ S\cap C'=\emptyset}} \psi_{C'}(\mathbf{x}_{C'}) \right) \left(\sum_{\mathbf{x}_S} \prod_{\substack{C\in\mathcal{C}:\\ S\cap C\neq\emptyset}} \psi_C(\mathbf{x}_C) \right). \end{aligned}$$

It comes

$$p(\mathbf{x}_S | \mathbf{x}_{\partial S}) = \frac{p(\mathbf{x}_S, \mathbf{x}_{\partial S})}{p(\mathbf{x}_{\partial S})} = \frac{\prod_{C: S\cap C\neq\emptyset} \psi_C(\mathbf{x}_C)}{\sum_{\widetilde{\mathbf{x}}_S} \prod_{C: S\cap C\neq\emptyset} \psi_C(\widetilde{\mathbf{x}}_C)}. \tag{5}$$

The final equalities in (4) and (5) are the same, thus proving that $p(\mathbf{x}_S|\mathbf{x}_{V\setminus S})$ and $p(\mathbf{x}_S|\mathbf{x}_{\partial S})$ are equal.

Problem 8 (Barber, p.99, Exercise 5.4)



1)

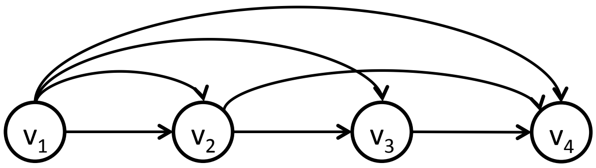
2) A simple linear chain for $p(\mathbf{h})$ can be easily seen from

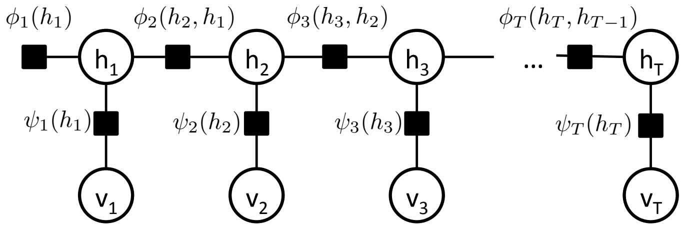
$$p(\mathbf{h}) = \sum_{\mathbf{v}} p(\mathbf{v}, \mathbf{h}) = p(h_1) \prod_{t=2}^{T} p(h_t | h_{t-1})$$

On the other hand, $p(\mathbf{v})$ is a fully connected cascade belief network because the marginal probability does not admit any decomposition. For example $T = 4$,

$$\begin{aligned} p(v_1, v_2, v_3, v_4) &= \sum_{h_1,h_2,h_3,h_4} p(v_1, v_2, v_3, v_4, h_1, h_2, h_3, h_4) \\ &= \sum_{h_4} p(v_4|h_4) \sum_{h_3} \Big(p(v_3, h_4|h_3) \sum_{h_2} \left(p(v_2, h_3|h_2) p(v_1, h_2) \right) \Big) \end{aligned}$$

We see that v_1, v_2, v_3, h_4 are all coupled.





3)

The factors are $\psi_t(h_t) = p(v_t|h_t)$, $\phi_1(h_1) = p(h_1)$ and $\phi_t(h_t, h_{t-1}) = p(h_t|h_{t-1})$ for $t \geq 2$.

4) Suppose our observation is $\hat{\mathbf{v}} = (\hat{v}_1, \ldots, \hat{v}_T)$. Since

$$p(\mathbf{h}|\mathbf{v} = \hat{\mathbf{v}}) \propto p(\mathbf{h}, \mathbf{v} = \hat{\mathbf{v}}),$$

we can use a sum-product algorithm to compute the marginal $p(h_t, \hat{\mathbf{v}})$ and then it is easy to obtain $p(h_t|\hat{\mathbf{v}}) = \frac{p(h_t,\hat{\mathbf{v}})}{\sum_{h_t} p(h_t,\hat{\mathbf{v}})}$. Recall that

$$\begin{aligned}
p(\hat{\mathbf{v}}, h_t) = \sum_{\mathbf{h}_{\sim t}} p(\hat{\mathbf{v}}, \mathbf{h}) &= \sum_{\mathbf{h}_{\sim t}} p(h_1)p(\hat{v}_1|h_1) \prod_{i=2}^{T} p(\hat{v}_i|h_i)p(h_i|h_{i-1}) \\
&= \sum_{\mathbf{h}_{\sim t}} \phi_1(h_1)\psi_1(h_1) \prod_{i=2}^{T} \psi_i(h_i)\phi_i(h_i, h_{i-1})
\end{aligned}$$

To compute the sum efficiently we define messages propagating from the two ends of the factor graph. For the forward propagation we define the factor-to-variable message

$$\mu_{\psi_i \to h_i}(h_i) = \psi(h_i), \quad \mu_{\phi_i \to h_i}(h_i) = \sum_{h_{i-1}} \phi_i(h_i, h_{i-1})\mu_{h_{i-1} \to \phi_i}(h_{i-1}) \text{ with } \phi_1(h_1, h_0) \triangleq \phi_1(h_1)$$

and variable-to-factor message

$$\mu_{h_i \to \phi_{i+1}}(h_i) = \mu_{\psi_i \to h_i}(h_i)\mu_{\phi_i \to h_i}(h_i)$$

We compute the messages in the order $(\mu_{\psi_1 \to h_1}, \mu_{\phi_1 \to h_1}) \to \mu_{h_1 \to \phi_2} \to (\mu_{\psi_2 \to h_2}, \mu_{\phi_2 \to h_2}) \to \mu_{h_2 \to \phi_3} \to \cdots \to (\mu_{\psi_t \to h_t}, \mu_{\phi_t \to h_t})$. So we have

$$\mu_{\phi_t \to h_t} = \sum_{h_1, \ldots, h_{t-1}} \psi_1(h_1)\psi_1(h_1) \prod_{i=2}^{t} \psi_i(h_i)\phi_i(h_i, h_{i-1})$$

It does not harm to continue the forward propagation up to $(\mu_{\psi_T \to h_T}, \mu_{\phi_T \to h_T})$ but here it is unneccessary. Next, we start the backward propagation with factor-to-variable message

$$\mu_{\phi_i \to h_{i-1}}(h_{i-1}) = \sum_{h_i} \phi_i(h_i, h_{i-1})\mu_{h_i \to \phi_i}(h_i)$$

and variable-to-factor message

$$\mu_{h_i \to \phi_i}(h_i) = \mu_{\psi_i \to h_i}(h_i)\mu_{\phi_{i+1} \to h_i}(h_i) \text{ with } \mu_{\phi_{T+1} \to h_T}(h_T) \triangleq 1$$

We proceed with $\mu_{\psi_T \to h_T} \to \mu_{h_T \to \phi_T} \to (\mu_{\psi_{T-1} \to h_{T-1}}, \mu_{\phi_T \to h_{T-1}}) \to \mu_{h_{T-1} \to \phi_{T-1}} \to \cdots \to \mu_{\phi_{t+1} \to h_t}$. So we have

$$\mu_{\phi_t \to h_t}(h_t) = \sum_{h_{t+1},\ldots,h_T} \prod_{i=t+1}^{T} \psi_i(h_i)\phi_i(h_i, h_{i-1})$$

and therefore

$$\begin{aligned} p(h_t, \hat{\mathbf{v}}) &= \mu_{\phi_t \to h_t}(h_t)\mu_{\psi_t \to h_t}(h_t)\mu_{\phi_{t+1} \to h_t}(h_t), \\ p(h_t|\hat{\mathbf{v}}) &= \frac{\mu_{\phi_t \to h_t}(h_t)\mu_{\psi_t \to h_t}(h_t)\mu_{\phi_{t+1} \to h_t}(h_t)}{\sum_{h_t} \mu_{\phi_t \to h_t}(h_t)\mu_{\psi_t \to h_t}(h_t)\mu_{\phi_{t+1} \to h_t}(h_t)}. \end{aligned}$$

5) Like the starting argument in the last question, we need to compute $\sum_{\mathbf{h}_{\sim t,t+1}} p(h_t, h_{t+1}, \hat{\mathbf{v}})$ where $\mathbf{h}_{\sim t,t+1}$ means h_t and h_{t+1} are excluded. So with the same message passing rules we obtain

$$p(h_t, h_{t+1}|\hat{\mathbf{v}}) \propto \mu_{\phi_t \to h_t}(h_t)\mu_{\psi_t \to h_t}(h_t)\phi_{t+1}(h_t, h_{t+1})\mu_{\phi_{t+2} \to h_{t+1}}(h_{t+1})\mu_{\psi_{t+1} \to h_{t+1}}(h_{t+1})$$

Problem 9 (Barber, p.98, Exercise 5.1)

The underlying undirected graph of a singly connected network with N nodes is a tree. We denote the tree with N nodes by $\mathcal{T}_N$. By definition it contains a leaf i which is connected to node j. The tree structure ensures the decomposition

$$Z = \sum_{\mathbf{x}_{\sim i}} \prod_{\substack{k \sim l \\ k \neq i \\ l \neq i}} \phi_{k,l}(x_k, x_l) \sum_{x_i} \phi_{i,j}(x_i, x_j).$$

where $\mathbf{x}_{\sim i}$ means x_i is excluded. So we can start the following recursion with $\mathcal{T}_N$.

1. Find a leaf i which is connected to node j.
2. Compute $\psi_{i,j}(x_j) = \sum_{x_i} \phi_{i,j}(x_i, x_j)$.
3. If node j has another neighbor node k,
3a. obtain $\mathcal{T}_{n-1}$ by removing node i and updating $\phi_{j,k}(x_j, x_k) \to \psi_{i,j}(x_j)\phi_{j,k}(x_j, x_k)$, and go to step 1 with $\mathcal{T}_{n-1}$;
3b. otherwise, there remain only node i and j, so we output $Z = \sum_{x_j} \psi_{i,j}(x_j)$.

The above algorithm ends with N iterations and therefore the time complexity is $O(N)$.

Problem 10 (Bishop, p.397 & 421, Exercise 8.16 & 8.17)

1) Given the observation $x_N = \hat{x}_N$, the initial message for β-recursion becomes

$$\mu_\beta(x_{N-1}) = \phi_{N-1,N}(x_{N-1}, \hat{x}_N).$$

Note that this initial message does not sum over x_N. The other message passing equations are unchanged. This message passing allows us to compute $p(x_n|x_N = \hat{x}_N)$.

2) Given the observation $x_3 = \hat{x}_3$, the algorithm suggests

$$p(x_2) = \frac{1}{Z}\mu_\alpha(x_2)\mu_\beta(x_2)$$

where

$$\begin{aligned}
&\mu_\beta(x_2) = \phi_{2,3}(x_2, \hat{x}_3)\mu_\beta(\hat{x}_3), \\
&Z = \sum_{x_2} \mu_\alpha(x_2)\mu_\beta(x_2) = \sum_{x_2} \mu_\alpha(x_2)\phi_{2,3}(x_2, \hat{x}_3)\mu_\beta(\hat{x}_3).
\end{aligned}$$

We can simplify the expression to

$$p(x_2) = \frac{\mu_\alpha(x_2)\phi_{2,3}(x_2, \hat{x}_3)}{\sum_{x_2} \mu_\alpha(x_2)\phi_{2,3}(x_2, \hat{x}_3)}.$$

Different x_5 will rescale $\mu_\beta(\hat{x}_3)$ but it changes nothing on $p(x_2)$. This aligns with the fact that $x_2 \perp\!\!\!\perp x_5|x_3$.