Exercises for the tutorials: 2 and 4.

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

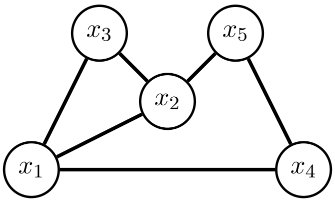
Exercise 1. *Visualising and analysing Gibbs distributions via undirected graphs*

We here consider the Gibbs distribution

$$p(x_1, \ldots, x_5) \propto \phi_{12}(x_1, x_2)\phi_{13}(x_1, x_3)\phi_{14}(x_1, x_4)\phi_{23}(x_2, x_3)\phi_{25}(x_2, x_5)\phi_{45}(x_4, x_5)$$

(a) *Visualise it as an undirected graph.*

Solution. We draw a node for each random variable x_i. There is an edge between two nodes if the corresponding variables co-occur in a factor.



(b) *What are the neighbours of x_3 in the graph?*

Solution. The neighbours are all the nodes for which there is a single connecting edge. Thus: $\text{ne}(x_3) = \{x_1, x_2\}$. (Note that sometimes, we may denote $\text{ne}(x_3)$ by ne_3.)

(c) *Do we have $x_3 \perp\!\!\!\perp x_4 \mid x_1, x_2$?*

Solution. Yes. The conditioning set $\{x_1, x_2\}$ equals ne_3, which is also the Markov blanket of x_3. This means that x_3 is conditionally independent of all the other variables given $\{x_1, x_2\}$, i.e. $x_3 \perp\!\!\!\perp x_4, x_5 \mid x_1, x_2$, which implies that $x_3 \perp\!\!\!\perp x_4 \mid x_1, x_2$. (One can also use graph separation to answer the question.)

(d) *What is the Markov blanket of x_4?*

Solution. The Markov blanket of a node in a undirected graphical model equals the set of its neighbours: $\text{MB}(x_4) = \text{ne}(x_4) = \text{ne}_4 = \{x_1, x_5\}$. This implies, for example, that $x_4 \perp\!\!\!\perp x_2, x_3 \mid x_1, x_5$.

(e) *On which minimal set of variables A do we need to condition to have $x_1 \perp\!\!\!\perp x_5 \mid A$?*

Solution. We first identify all trails from x_1 to x_5. There are three such trails: (x_1, x_2, x_5), (x_1, x_3, x_2, x_5), and (x_1, x_4, x_5). Conditioning on x_2 blocks the first two trails, conditioning on x_4 blocks the last. We thus have: $x_1 \perp\!\!\!\perp x_5 \mid x_2, x_4$, so that $A = \{x_2, x_4\}$.

Exercise 2. ***Factorisation and independencies for undirected graphical models***

Consider the undirected graphical model defined by the graph in Figure 1.

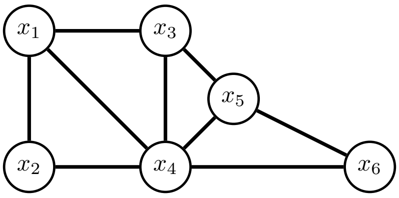


Figure 1: Graph for Exercise 2

(a) *What is the set of Gibbs distributions that is induced by the graph?*

Solution. The graph in Figure 1 has four maximal cliques:

$$(x_1, x_2, x_4) \quad (x_1, x_3, x_4) \quad (x_3, x_4, x_5) \quad (x_4, x_5, x_6)$$

The Gibbs distributions are thus

$$p(x_1, \dots, x_6) \propto \phi_1(x_1, x_2, x_4)\phi_2(x_1, x_3, x_4)\phi_3(x_3, x_4, x_5)\phi_4(x_4, x_5, x_6)$$

(b) *Let p be a pdf that factorises according to the graph. Does $p(x_3|x_2, x_4) = p(x_3|x_4)$ hold?*

Solution. $p(x_3|x_2, x_4) = p(x_3|x_4)$ means that $x_3 \perp\!\!\!\perp x_2 \mid x_4$. We can use the graph to check whether this generally holds for pdfs that factorise according to the graph. There are multiple trails from x_3 to x_2, including the trail (x_3, x_1, x_2), which is not blocked by x_4. From the graph, we thus cannot conclude that $x_3 \perp\!\!\!\perp x_2 \mid x_4$, and $p(x_3|x_2, x_4) = p(x_3|x_4)$ will generally not hold (the relation may hold for some carefully defined factors ϕ_i).

(c) *Explain why $x_2 \perp\!\!\!\perp x_5 \mid x_1, x_3, x_4, x_6$ holds for all distributions that factorise over the graph.*

Solution. Distributions that factorise over the graph satisfy the pairwise Markov property. Since x_2 and x_5 are not neighbours, and x_1, x_3, x_4, x_6 are the remaining nodes in the graph, the independence relation follows from the pairwise Markov property.

(d) *Assume you would like to approximate $\mathbb{E}(x_1 x_2 x_5 \mid x_3, x_4)$, i.e. the expected value of the product of x_1, x_2, and x_5 given x_3 and x_4, with a sample average. Do you need to have joint observations for all five variables $x_1, \dots, x_5$?*

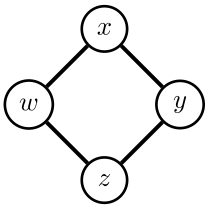
Solution. In the graph, all trails from $\{x_1, x_2\}$ to x_5 are blocked by $\{x_3, x_4\}$, so that $x_1, x_2 \perp\!\!\!\perp x_5 \mid x_3, x_4$. We thus have

$$\mathbb{E}(x_1 x_2 x_5 \mid x_3, x_4) = \mathbb{E}(x_1 x_2 \mid x_3, x_4)\mathbb{E}(x_5 \mid x_3, x_4).$$

Hence, we only need joint observations of (x_1, x_2, x_3, x_4) and (x_3, x_4, x_5). Variables (x_1, x_2) and x_5 do not need to be jointly measured.

Exercise 3. *Factorisation and independencies for undirected graphical models*

Consider the undirected graphical model defined by the following graph, sometimes called a diamond configuration.



(a) *How do the pdfs/pmfs of the undirected graphical model factorise?*

Solution. The maximal cliques are (x, w), (w, z), (z, y) and (x, y). The undirected graphical model thus consists of pdfs/pmfs that factorise as follows

$$p(x, w, z, y) \propto \phi_1(x, w)\phi_2(w, z)\phi_3(z, y)\phi_4(x, y) \tag{S.1}$$

(b) *List all independencies that hold for the undirected graphical model.*

Solution. We can generate the independencies by conditioning on progressively larger sets. Since there is a trail between any two nodes, there are no unconditional independencies. If we condition on a single variable, there is still a trail that connects the remaining ones. Let us thus consider the case where we condition on two nodes. By graph separation, we have

$$w \perp\!\!\!\perp y \mid x, z \qquad x \perp\!\!\!\perp z \mid w, y \tag{S.2}$$

These are all the independencies that hold for the model, since conditioning on three nodes does not lead to any independencies in a model with four variables.

Exercise 4. *Factorisation from the Markov blankets I*

Assume you know the following Markov blankets for all variables $x_1, \ldots, x_4, y_1, \ldots y_4$ of a pdf or pmf $p(x_1, \ldots, x_4, y_1, \ldots, y_4)$.

$$MB(x_1) = \{x_2, y_1\} \quad MB(x_2) = \{x_1, x_3, y_2\} \quad MB(x_3) = \{x_2, x_4, y_3\} \quad MB(x_4) = \{x_3, y_4\} \tag{1}$$

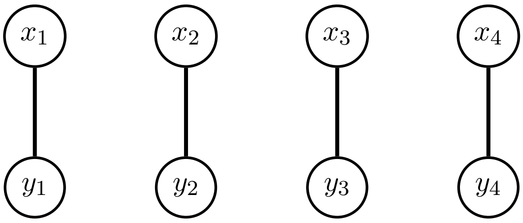
$$MB(y_1) = \{x_1\} \quad MB(y_2) = \{x_2\} \quad MB(y_3) = \{x_3\} \quad MB(y_4) = \{x_4\} \tag{2}$$

Assuming that p is positive for all possible values of its variables, how does p factorise?

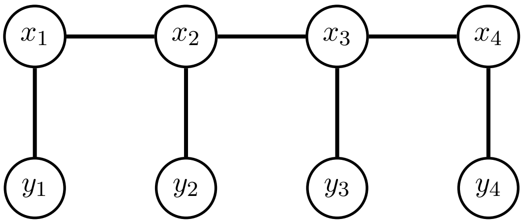
Solution. In undirected graphical models, the Markov blanket for a variable is the same as the set of its neighbours. Hence, when we are given all Markov blankets we know what local Markov property p must satisfy. For positive distributions we have an equivalence between p satisfying the local Markov property and p factorising over the graph. Hence, to specify the factorisation of p it suffices to construct the undirected graph H based on the Markov blankets and then read out the factorisation.

We need to build a graph where the neighbours of each variable equals the indicated Markov blanket. This can be easily done by starting with an empty graph and connecting each variable to the variables in its Markov blanket.

We see that each y_i is only connected to x_i. Including those Markov blankets we get the following graph:



Connecting the x_i to their neighbours according to the Markov blanket thus gives:



The graph has maximal cliques of size two, namely the $x_i - y_i$ for $i = 1, \ldots, 4$, and the $x_i - x_{i+1}$ for $i = 1, \ldots, 3$. Given the equivalence between the local Markov property and factorisation for positive distributions, we know that p must factorise as

$$p(x_1, \ldots, x_4, y_1, \ldots, y_4) = \frac{1}{Z} \prod_{i=1}^{3} m_i(x_i, x_{i+1}) \prod_{i=1}^{4} g_i(x_i, y_i), \tag{S.3}$$

where $m_i(x_i, x_{i+1}) > 0$, $g(x_i, y_i) > 0$ are positive factors (potential functions).

The graphical model corresponds to an undirected version of a hidden Markov model where the x_i are the unobserved (latent, hidden) variables and the y_i are the observed ones. Note that the x_i form a Markov chain.

Exercise 5. *Factorisation from the Markov blankets II*

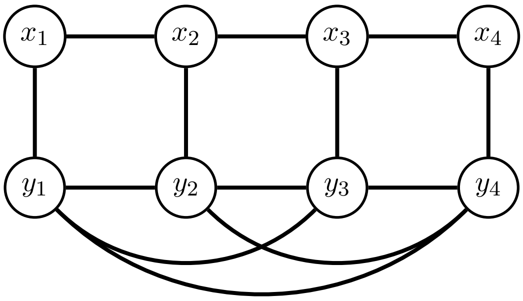
We consider the same setup as in Exercise 4 but we now assume that we do not know all Markov blankets but only

$$MB(x_1) = \{x_2, y_1\} \qquad MB(x_2) = \{x_1, x_3, y_2\} \qquad MB(x_3) = \{x_2, x_4, y_3\} \qquad MB(x_4) = \{x_3, y_4\} \tag{3}$$

Without inserting more independencies than those specified by the Markov blankets, draw the graph over which p factorises and state the factorisation. (Again assume that p is positive for all possible values of its variables).

Solution. We take the same approach as in Exercise 4. In particular, the Markov blankets of a variable are its neighbours in the graph. But since we are not given all Markov blankets and are not allowed to insert additional independencies, we must assume that each y_i is connected to all the other $y's$. For example, if we didn't connect y_1 and y_4 we would assert the additional independency $y_1 \perp\!\!\!\perp y_4 \mid x_1, x_2, x_3, x_4, y_2, y_3$.

We thus have a graph as follows:



The factorisation thus is

$$p(x_1, \dots, x_4, y_1, \dots, y_4) = \frac{1}{Z} g(y_1, \dots, y_4) \prod_{i=1}^{3} m_i(x_i, x_{i+1}) \prod_{i=1}^{4} g_i(x_i, y_i), \tag{S.4}$$

where the $m_i(x_i, x_{i+1})$, $g_i(x_i, y_i)$ and $g(y_1, \dots, y_4)$ are positive factors. Compared to the factorisation in Exercise 4, we still have the Markov structure for the x_i, but only a single factor for (y_1, y_2, y_3, y_4) to avoid inserting independencies beyond those specified by the given Markov blankets.

Exercise 6. *Undirected graphical model with pairwise potentials*

We here consider Gibbs distributions where the factors only depend on two variables at a time. The probability density or mass functions over d random variables $x_1, \dots, x_d$ then take the form

$$p(x_1, \dots, x_d) \propto \prod_{i \leq j} \phi_{ij}(x_i, x_j)$$

Such models are sometimes called pairwise Markov networks.

(a) *Let $p(x_1, \dots, x_d) \propto \exp\left(-\frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{x}\right)$ where $\mathbf{A}$ is symmetric and $\mathbf{x} = (x_1, \dots, x_d)^\top$. What are the corresponding factors ϕ_{ij} for $i \leq j$?*

Solution. Denote the (i, j)-th element of $\mathbf{A}$ by a_{ij}. We have

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} = \sum_{ij} a_{ij} x_i x_j \tag{S.5}$$

$$= \sum_{i<j} 2a_{ij} x_i x_j + \sum_i a_{ii} x_i^2 \tag{S.6}$$

where the second line follows from $\mathbf{A}^\top = \mathbf{A}$. Hence,

$$-\frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{x} = -\frac{1}{2} \sum_{i<j} 2a_{ij} x_i x_j - \frac{1}{2} \sum_i a_{ii} x_i^2 - \sum_i b_i x_i \tag{S.7}$$

so that

$$\phi_{ij}(x_i, x_j) = \begin{cases} \exp\left(-a_{ij} x_i x_j\right) & \text{if } i < j \\ \exp\left(-\frac{1}{2} a_{ii} x_i^2 - b_i x_i\right) & \text{if } i = j \end{cases} \tag{S.8}$$

For $\mathbf{x} \in \mathbb{R}^d$, the distribution is a Gaussian with $\mathbf{A}$ equal to the inverse covariance matrix. For binary $\mathbf{x}$, the model is known as Ising model or Boltzmann machine. For $x_i \in \{-1, 1\}$,

$x_i^2 = 1$ for all i, so that the a_{ii} are constants that can be absorbed into the normalisation constant. This means that for $x_i \in \{-1, 1\}$, we can work with matrices $\mathbf{A}$ that have zeros on the diagonal.

(b) *For $p(x_1, \ldots, x_d) \propto \exp\left(-\frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{x}\right)$, show that $x_i \perp\!\!\!\perp x_j \mid \{x_1, \ldots, x_d\} \setminus \{x_i, x_j\}$ if the (i,j)-th element of $\mathbf{A}$ is zero.*

Solution. The previous question showed that we can write $p(x_1, \ldots, x_d) \propto \prod_{i \leq j} \phi_{ij}(x_i, x_j)$ with potentials as in Equation (S.8). Consider two variables x_i and x_j for fixed (i, j). They only appear in the factorisation via the potential ϕ_{ij}. If $a_{ij} = 0$, the factor ϕ_{ij} becomes a constant, and no other factor contains x_i and x_j, which means that there is no edge between x_i and x_j if $a_{ij} = 0$. By the pairwise Markov property it then follows that $x_i \perp\!\!\!\perp x_j \mid \{x_1, \ldots, x_d\} \setminus \{x_i, x_j\}$.

Exercise 7. *Restricted Boltzmann machine (based on Barber Exercise 4.4)*

The restricted Boltzmann machine is an undirected graphical model for binary variables $\mathbf{v} = (v_1, \ldots, v_n)^\top$ and $\mathbf{h} = (h_1, \ldots, h_m)^\top$ with a probability mass function equal to

$$p(\mathbf{v}, \mathbf{h}) \propto \exp\left(\mathbf{v}^\top \mathbf{W}\mathbf{h} + \mathbf{a}^\top \mathbf{v} + \mathbf{b}^\top \mathbf{h}\right), \tag{4}$$

where $\mathbf{W}$ is a $n \times m$ matrix. Both the v_i and h_i take values in $\{0, 1\}$. The v_i are called the "visibles" variables since they are assumed to be observed while the h_i are the hidden variables since it is assumed that we cannot measure them.

(a) *Use graph separation to show that the joint conditional $p(\mathbf{h}|\mathbf{v})$ factorises as*

$$p(\mathbf{h}|\mathbf{v}) = \prod_{i=1}^{m} p(h_i|\mathbf{v}).$$

Solution. Figure 2 on the left shows the undirected graph for $p(\mathbf{v}, \mathbf{h})$ with $n = 3, m = 2$. We note that the graph is bi-partite: there are only direct connections between the h_i and the v_i. Conditioning on $\mathbf{v}$ thus blocks all trails between the h_i (graph on the right). This means that the h_i are independent from each other given $\mathbf{v}$ so that

$$p(\mathbf{h}|\mathbf{v}) = \prod_{i=1}^{m} p(h_i|\mathbf{v}).$$

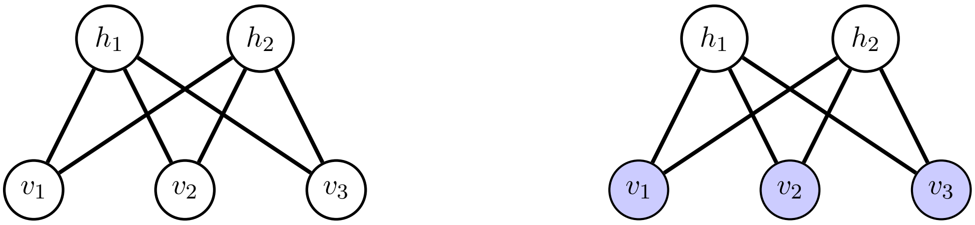


Figure 2: Left: Graph for $p(\mathbf{v}, \mathbf{h})$. Right: Graph for $p(\mathbf{h}|\mathbf{v})$

(b) Show that

$$p(h_i = 1|\mathbf{v}) = \frac{1}{1 + \exp\left(-b_i - \sum_j W_{ji} v_j\right)} \tag{5}$$

where W_{ji} is the (ji)-th element of $\mathbf{W}$, so that $\sum_j W_{ji} v_j$ is the inner product (scalar product) between the i-th column of $\mathbf{W}$ and $\mathbf{v}$.

Solution. For the conditional pmf $p(h_i|\mathbf{v})$ any quantity that does not depend on h_i can be considered to be part of the normalisation constant. A general strategy is to first work out $p(h_i|\mathbf{v})$ up to the normalisation constant and then to normalise it afterwards.

We begin with $p(\mathbf{h}|\mathbf{v})$:

$$\begin{aligned}
p(\mathbf{h}|\mathbf{v}) &= \frac{p(\mathbf{h}, \mathbf{v})}{p(\mathbf{v})} && \text{(S.9)}\\
&\propto p(\mathbf{h}, \mathbf{v}) && \text{(S.10)}\\
&\propto \exp\left(\mathbf{v}^\top \mathbf{W} \mathbf{h} + \mathbf{a}^\top \mathbf{v} + \mathbf{b}^\top \mathbf{h}\right) && \text{(S.11)}\\
&\propto \exp\left(\mathbf{v}^\top \mathbf{W} \mathbf{h} + \mathbf{b}^\top \mathbf{h}\right) && \text{(S.12)}\\
&\propto \exp\left(\sum_i \sum_j v_j W_{ji} h_i + \sum_i b_i h_i\right) && \text{(S.13)}
\end{aligned}$$

As we are interested in $p(h_i|\mathbf{v})$ for a fixed i, we can drop all the terms not depending on that h_i, so that

$$p(h_i|\mathbf{v}) \propto \exp\left(\sum_j v_j W_{ji} h_i + b_i h_i\right) \tag{S.14}$$

Since h_i only takes two values, 0 and 1, normalisation is here straightforward. Call the unnormalised pmf $\tilde{p}(h_i|\mathbf{v})$,

$$\tilde{p}(h_i|\mathbf{v}) = \exp\left(\sum_j v_j W_{ji} h_i + b_i h_i\right). \tag{S.15}$$

We then have

$$\begin{aligned}
p(h_i|\mathbf{v}) &= \frac{\tilde{p}(h_i|\mathbf{v})}{\tilde{p}(h_i = 0|\mathbf{v}) + \tilde{p}(h_i = 1|\mathbf{v})} && \text{(S.16)}\\
&= \frac{\tilde{p}(h_i|\mathbf{v})}{1 + \exp\left(\sum_j v_j W_{ji} + b_i\right)} && \text{(S.17)}\\
&= \frac{\exp\left(\sum_j v_j W_{ji} h_i + b_i h_i\right)}{1 + \exp\left(\sum_j v_j W_{ji} + b_i\right)}, && \text{(S.18)}
\end{aligned}$$

so that

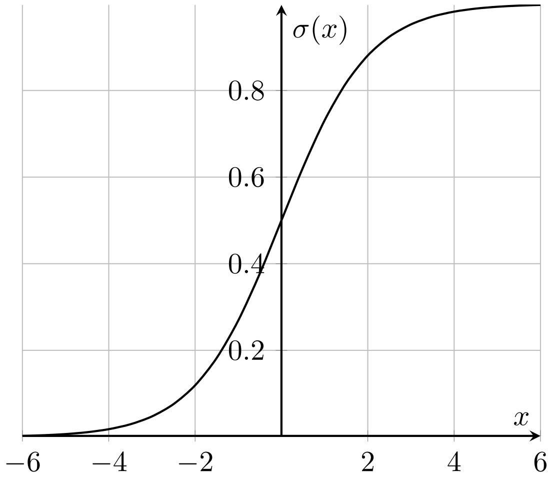
$$\begin{aligned}
p(h_i = 1|\mathbf{v}) &= \frac{\exp\left(\sum_j v_j W_{ji} + b_i\right)}{1 + \exp\left(\sum_j v_j W_{ji} + b_i\right)} && \text{(S.19)}\\
&= \frac{1}{1 + \exp\left(-\sum_j v_j W_{ji} - b_i\right)}. && \text{(S.20)}
\end{aligned}$$

The probability $p(h = 0|\mathbf{v})$ equals $1 - p(h_i = 1|\mathbf{v})$, which is

$$p(h_i = 0|\mathbf{v}) = \frac{1 + \exp\left(\sum_j v_j W_{ji} + b_i\right)}{1 + \exp\left(\sum_j v_j W_{ji} + b_i\right)} - \frac{\exp\left(\sum_j v_j W_{ji} + b_i\right)}{1 + \exp\left(\sum_j v_j W_{ji} + b_i\right)} \tag{S.21}$$

$$= \frac{1}{1 + \exp\left(\sum_j W_{ji} v_j + b_i\right)} \tag{S.22}$$

The function $x \mapsto 1/(1 + \exp(-x))$ is called the logistic function. It is a sigmoid function and is thus sometimes denoted by $\sigma(x)$. For other versions of the sigmoid function, see https://en.wikipedia.org/wiki/Sigmoid_function.



With that notation, we have

$$p(h_i = 1|\mathbf{v}) = \sigma\left(\sum_j W_{ji} v_j + b_i\right).$$

(c) Use a symmetry argument to show that

$$p(\mathbf{v}|\mathbf{h}) = \prod_i p(v_i|\mathbf{h}) \quad \textit{and} \quad p(v_i = 1|\mathbf{h}) = \frac{1}{1 + \exp\left(-a_i - \sum_j W_{ij} h_j\right)}$$

Solution. Since $\mathbf{v}^\top \mathbf{W} \mathbf{h}$ is a scalar we have $(\mathbf{v}^\top \mathbf{W} \mathbf{h})^\top = \mathbf{h}^\top \mathbf{W}^\top \mathbf{v} = \mathbf{v}^\top \mathbf{W} \mathbf{h}$, so that

$$p(\mathbf{v}, \mathbf{h}) \propto \exp\left(\mathbf{v}^\top \mathbf{W} \mathbf{h} + \mathbf{a}^\top \mathbf{v} + \mathbf{b}^\top \mathbf{h}\right) \tag{S.23}$$

$$\propto \exp\left(\mathbf{h}^\top \mathbf{W}^\top \mathbf{v} + \mathbf{b}^\top \mathbf{h} + \mathbf{a}^\top \mathbf{v}\right). \tag{S.24}$$

To derive the result, we note that $\mathbf{v}$ and a now take the place of $\mathbf{h}$ and $\mathbf{b}$ from before, and that we now have $\mathbf{W}^\top$ rather than $\mathbf{W}$. In Equation (5), we thus replace h_i with v_i, b_i with a_i, and W_{ji} with W_{ij} to obtain $p(v_i = 1|\mathbf{h})$. In terms of the sigmoid function, we have

$$p(v_i = 1|\mathbf{h}) = \sigma\left(\sum_j W_{ij} h_j + a_i\right).$$

Note that while $p(\mathbf{v}|\mathbf{h})$ factorises, the marginal $p(\mathbf{v})$ does generally not. The marginal $p(\mathbf{v})$ can here be obtained in closed form up to its normalisation constant.

$$\begin{aligned}
p(\mathbf{v}) &= \sum_{\mathbf{h}\in\{0,1\}^m} p(\mathbf{v},\mathbf{h}) && \text{(S.25)}\\
&= \frac{1}{Z}\sum_{\mathbf{h}\in\{0,1\}^m} \exp\left(\mathbf{v}^\top\mathbf{W}\mathbf{h} + \mathbf{a}^\top\mathbf{v} + \mathbf{b}^\top\mathbf{h}\right) && \text{(S.26)}\\
&= \frac{1}{Z}\sum_{\mathbf{h}\in\{0,1\}^m} \exp\left(\sum_{ij} v_i h_j W_{ij} + \sum_i a_i v_i + \sum_j b_j h_j\right) && \text{(S.27)}\\
&= \frac{1}{Z}\sum_{\mathbf{h}\in\{0,1\}^m} \exp\left(\sum_{j=1}^m h_j\left[\sum_i v_i W_{ij} + b_j\right] + \sum_i a_i v_i\right) && \text{(S.28)}\\
&= \frac{1}{Z}\sum_{\mathbf{h}\in\{0,1\}^m} \prod_{j=1}^m \exp\left(h_j\left[\sum_i v_i W_{ij} + b_j\right]\right)\exp\left(\sum_i a_i v_i\right) && \text{(S.29)}\\
&= \frac{1}{Z}\exp\left(\sum_i a_i v_i\right)\sum_{\mathbf{h}\in\{0,1\}^m} \prod_{j=1}^m \exp\left(h_j\left[\sum_i v_i W_{ij} + b_j\right]\right) && \text{(S.30)}\\
&= \frac{1}{Z}\exp\left(\sum_i a_i v_i\right)\sum_{h_1,\ldots,h_m} \prod_{j=1}^m \exp\left(h_j\left[\sum_i v_i W_{ij} + b_j\right]\right) && \text{(S.31)}
\end{aligned}$$

Importantly, each term in the product only depends on a single h_j, so that by sequentially applying the distributive law, we have

$$\begin{aligned}
\sum_{h_1,\ldots,h_m}\prod_{j=1}^m \exp\left(h_j\left[\sum_i v_i W_{ij} + b_j\right]\right) &= \left[\sum_{h_1,\ldots,h_{m-1}}\prod_{j=1}^{m-1} \exp\left(h_j\left[\sum_i v_i W_{ij} + b_j\right]\right)\right]\cdot\\
&\quad\sum_{h_m}\exp\left(h_m\left[\sum_i v_i W_{im} + b_m\right]\right) && \text{(S.32)}\\
&= \ldots\\
&= \prod_{j=1}^m\left[\sum_{h_j}\exp\left(h_j\left[\sum_i v_i W_{ij} + b_j\right]\right)\right] && \text{(S.33)}
\end{aligned}$$

Since $h_j \in \{0,1\}$, we obtain

$$\sum_{h_j}\exp\left(h_j\left[\sum_i v_i W_{ij} + b_j\right]\right) = 1 + \exp\left(\sum_i v_i W_{ij} + b_j\right) \tag{S.34}$$

and thus

$$p(\mathbf{v}) = \frac{1}{Z}\exp\left(\sum_i a_i v_i\right)\prod_{j=1}^m\left[1 + \exp\left(\sum_i v_i W_{ij} + b_j\right)\right]. \tag{S.35}$$

Note that in the derivation of $p(\mathbf{v})$ we have not used the assumption that the visibles v_i are binary. The same expression would thus obtained if the visibles were defined in another space, e.g. the real numbers.

While $p(\mathbf{v})$ is written as a product, $p(\mathbf{v})$ does not factorise into terms that depend on subsets of the v_i. On the contrary, all v_i are present in all factors. Since $p(\mathbf{v})$ does not factorise, computing the normalising Z is expensive. For binary visibles $v_i \in \{0,1\}$, Z equals

$$Z = \sum_{\mathbf{v}\in\{0,1\}^n} \exp\left(\sum_i a_i v_i\right) \prod_{j=1}^{m} \left[1 + \exp\left(\sum_i v_i W_{ij} + b_j\right)\right] \tag{S.36}$$

where we have to sum over all 2^n configurations of the visibles $\mathbf{v}$. This is computationally expensive, or even prohibitive if n is large ($2^{20} = 1048576$, $2^{30} > 10^9$). Note that different values of a_i, b_i, W_{ij} yield different values of Z. (This is a reason why Z is called the partition *function* when the a_i, b_i, W_{ij} are free parameters.)

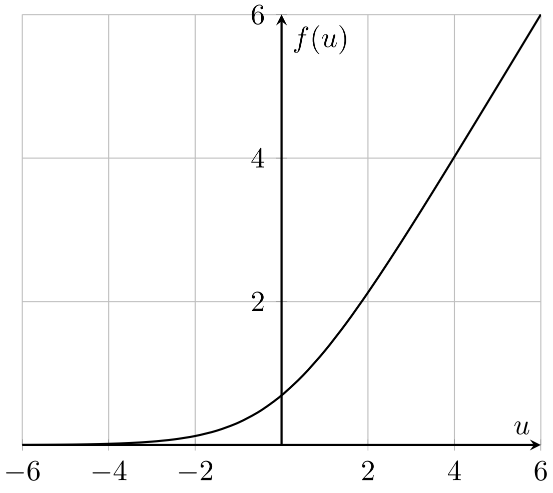
It is instructive to write $p(\mathbf{v})$ in the log-domain,

$$\log p(\mathbf{v}) = \log Z + \sum_{i=1}^{n} a_i v_i + \sum_{j=1}^{m} \log\left[1 + \exp\left(\sum_i v_i W_{ij} + b_j\right)\right], \tag{S.37}$$

and to introduce the nonlinearity $f(u)$,

$$f(u) = \log\left[1 + \exp(u)\right], \tag{S.38}$$

which is called the softplus function and plotted below. The softplus function is a smooth approximation of $\max(0, u)$, see e.g. https://en.wikipedia.org/wiki/Rectifier_(neural_networks)

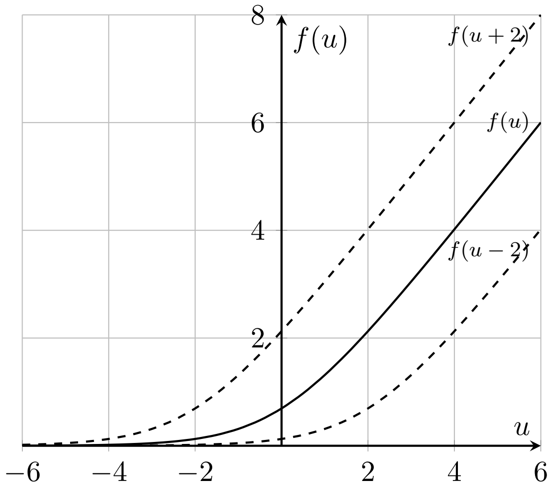


With the softplus function $f(u)$, we can write $\log p(\mathbf{v})$ as

$$\log p(\mathbf{v}) = \log Z + \sum_{i=1}^{n} a_i v_i + \sum_{j=1}^{m} f\left(\sum_i v_i W_{ij} + b_j\right). \tag{S.39}$$

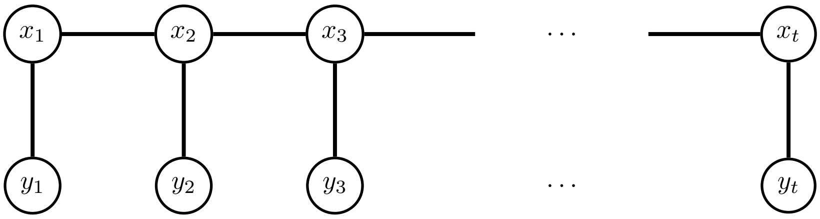
The parameter b_j plays the role of a threshold as shown in the figure below. The terms $f\left(\sum_i v_i W_{ij} + b_j\right)$ can be interpreted in terms of feature detection. The sum $\sum_i v_i W_{ij}$ is the inner product between $\mathbf{v}$ and the j-th column of $\mathbf{W}$, and the inner product is largest if $\mathbf{v}$ equals the j-th column. We can thus consider the columns of $\mathbf{W}$ to be feature-templates, and the $f\left(\sum_i v_i W_{ij} + b_j\right)$ a way to measure how much of each feature is present in $\mathbf{v}$.

Further, $\sum_i v_i W_{ij} + b_j$ is also the input to the sigmoid function when computing $p(h_j = 1|\mathbf{v})$. Thus, the conditional probability for h_j to be one, i.e. "active", can be considered to be an indicator of the presence of the j-th feature (j-th column of $\mathbf{W}$) in the input $\mathbf{v}$. If v is such that $\sum_i v_i W_{ij} + b_j$ is large for many j, i.e. if many features are detected, then $f\left(\sum_i v_i W_{ij} + b_j\right)$ will be non-zero for many j, and $\log p(\mathbf{v})$ will be large.



Exercise 8. ***Hidden Markov models and change of measure***

Consider the following undirected graph for a hidden Markov model where the y_i correspond to observed (visible) variables and the x_i to unobserved (hidden/latent) variables.



The graph implies the following factorisation

$$p(x_1, \ldots, x_t, y_1, \ldots, y_t) \propto \phi_1^y(x_1, y_1) \prod_{i=2}^{t} \phi_i^x(x_{i-1}, x_i) \phi_i^y(x_i, y_i), \tag{6}$$

where the ϕ_i^x and ϕ_i^y are non-negative factors.

Let us consider the situation where $\prod_{i=2}^t \phi_i^x(x_{i-1}, x_i)$ equals

$$f(\mathbf{x}) = \prod_{i=2}^{t} \phi_i^x(x_{i-1}, x_i) = f_1(x_1) \prod_{i=2}^{t} f_i(x_i|x_{i-1}), \tag{7}$$

with $\mathbf{x} = (x_1, \ldots, x_t)$ and where the f_i are (conditional) pdfs. We thus have

$$p(x_1, \ldots, x_t, y_1, \ldots, y_t) \propto f_1(x_1) \prod_{i=2}^{t} f_i(x_i|x_{i-1}) \prod_{i=1}^{t} \phi_i^y(x_i, y_i). \tag{8}$$

(a) *Provide a factorised expression for $p(x_1, \ldots, x_t|y_1, \ldots, y_t)$*

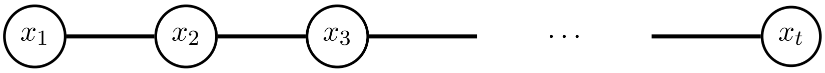
Solution. For fixed (observed) values of the y_i, $p(x_1, \ldots, x_t | y_1, \ldots, y_t)$ factorises as

$$p(x_1, \ldots, x_t | y_1, \ldots, y_t) \propto f_1(x_1) g_1(x_1) \prod_{i=2}^{t} f_i(x_i | x_{i-1}) g_i(x_i). \tag{S.40}$$

where $g_i(x_i)$ is $\phi_i^y(x_i, y_i)$ for a fixed value of y_i.

(b) *Draw the undirected graph for $p(x_1, \ldots, x_t | y_1, \ldots, y_t)$*

Solution. Conditioning corresponds to removing nodes from an undirected graph. We thus have the following Markov chain for $p(x_1, \ldots, x_t | y_1, \ldots, y_t)$.



(c) *Show that if $\phi_i^y(x_i, y_i)$ equals the conditional pdf of y_i given x_i, i.e. $p(y_i | x_i)$, the marginal $p(x_1, \ldots, x_t)$, obtained by integrating out $y_1, \ldots, y_t$ from (8), equals $f(\mathbf{x})$.*

Solution. In this setting all factors in (8) are conditional pdfs and we are dealing with a directed graphical model that factorises as

$$p(x_1, \ldots, x_t, y_1, \ldots, y_t) = f_1(x_1) \prod_{i=2}^{t} f_i(x_i | x_{i-1}) \prod_{i=1}^{t} p(y_i | x_i). \tag{S.41}$$

By integrating over the y_i, we have

$$\begin{aligned}
p(x_1, \ldots, x_t) &= \int p(x_1, \ldots, x_t, y_1, \ldots, y_t) \mathrm{d}y_1 \ldots \mathrm{d}y_t && \text{(S.42)} \\
&= f_1(x_1) \prod_{i=2}^{t} f_i(x_i | x_{i-1}) \int \prod_{i=1}^{t} p(y_i | x_i) \mathrm{d}y_1 \ldots \mathrm{d}y_t && \text{(S.43)} \\
&= f_1(x_1) \prod_{i=2}^{t} f_i(x_i | x_{i-1}) \prod_{i=1}^{t} \underbrace{\int p(y_i | x_i) \mathrm{d}y_i}_{1} && \text{(S.44)} \\
&= f_1(x_1) \prod_{i=2}^{t} f_i(x_i | x_{i-1}) && \text{(S.45)} \\
&= f(\mathbf{x}) && \text{(S.46)}
\end{aligned}$$

(d) *Compute the normalising constant for $p(x_1, \ldots, x_t | y_1, \ldots, y_t)$ and express it as an expectation over $f(\mathbf{x})$.*

Solution. With

$$p(x_1, \ldots, x_t, y_1, \ldots, y_t) \propto f_1(x_1) \prod_{i=2}^{t} f_i(x_i | x_{i-1}) \prod_{i=1}^{t} \phi_i^y(x_i, y_i). \tag{S.47}$$

The normalising constant is given by

$$Z = \int f_1(x_1) \prod_{2=1}^{t} f_i(x_i|x_{i-1}) \prod_{i=1}^{t} g_i(x_i) \mathrm{d}x_1 \ldots \mathrm{d}x_t \tag{S.48}$$

$$= \mathbb{E}_f \left[\prod_{i=1}^{t} g_i(x_i) \right] \tag{S.49}$$

Since we can use ancestral sampling to sample from f, the above expectation can be easily computed via sampling.

(e) *Express the expectation of a test function $h(\mathbf{x})$ with respect to $p(x_1, \ldots, x_t|y_1, \ldots, y_t)$ as a reweighted expectation with respect to $f(\mathbf{x})$.*

Solution. By definition, the expectation over a test function $h(\mathbf{x})$ is

$$\mathbb{E}_{p(x_1,\ldots,x_t|y_1,\ldots,y_t)}[h(\mathbf{x})] = \frac{1}{Z} \int h(\mathbf{x}) f_1(x_1) \prod_{i=2}^{t} f(x_i|x_{i-1}) \prod_{i=1}^{t} g_i(x_i) \mathrm{d}x_1 \ldots \mathrm{d}x_t \tag{S.50}$$

$$= \frac{\mathbb{E}_f \left[h(\mathbf{x}) \prod_i g_i(x_i) \right]}{\mathbb{E}_f \left[\prod_i g_i(x_i) \right]} \tag{S.51}$$

Both the numerator and denominator can be approximated using samples from f.

Since the $g_i(x_i) = \phi_i^y(x_i, y_i)$ involve the observed variables y_i, this has a nice interpretation: We can think we have two models for $\mathbf{x}$: $f(\mathbf{x})$ that does not involve the observations and $p(x_1, \ldots, x_t|y_1, \ldots, y_t)$ that does. Note, however, that unless $\phi_i^y(x_i, y_i)$ is the conditional pdf $p(y_i|x_i)$, $f(\mathbf{x})$ is *not* the marginal $p(x_1, \ldots, x_t)$ that you would obtain by integrating out the y's from the joint model . We can thus generally think it is a base distribution that got "enhanced" by a change of measure in our expression for $p(x_1, \ldots, x_t|y_1, \ldots, y_t)$. If $\phi_i^y(x_i, y_i)$ is the conditional pdf $p(y_i|x_i)$, the change of measure corresponds to going from the prior to the posterior by multiplication with the likelihood (the terms g_i).

From the expression for the expectation, we can see that the "enhancing" leads to a corresponding introduction of weights in the expectation that depend via g_i on the observations. This can be particularly well seen when we approximate the expectation as a sample average over n samples $\mathbf{x}^{(k)} \sim f(\mathbf{x})$:

$$\mathbb{E}_{p(x_1,\ldots,x_t|y_1,\ldots,y_t)}[h(\mathbf{x})] \approx \sum_{k=1}^{n} W^{(k)} h(\mathbf{x}^{(k)}) \tag{S.52}$$

$$W^{(k)} = \frac{w^{(k)}}{\sum_{k=1}^{n} w^{(k)}} \tag{S.53}$$

$$w^{(k)} = \prod_i g_i(x_i^{(k)}) \tag{S.54}$$

where $x_i^{(k)}$ is the i-th dimension of the vector $\mathbf{x}^{(k)}$.