## Exercise 1. Restricted Boltzmann machine (based on Barber Exercise 4.4)

The restricted Boltzmann machine is an undirected graphical model for binary variables $\mathbf{v}=\left(v_{1}, \ldots, v_{n}\right)^{\top}$ and $\mathbf{h}=\left(h_{1}, \ldots, h_{m}\right)^{\top}$ with a probability mass function equal to

$$
\begin{equation*}
p(\mathbf{v}, \mathbf{h}) \propto \exp \left(\mathbf{v}^{\top} \mathbf{W} \mathbf{h}+\mathbf{a}^{\top} \mathbf{v}+\mathbf{b}^{\top} \mathbf{h}\right) \tag{1}
\end{equation*}
$$

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 $p(\mathbf{h} \mid \mathbf{v})$ factorises as

$$
p(\mathbf{h} \mid \mathbf{v})=\prod_{i=1}^{m} p\left(h_{i} \mid \mathbf{v}\right)
$$

Solution. Figure 1 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

$$
p(\mathbf{h} \mid \mathbf{v})=\prod_{i=1}^{m} p\left(h_{i} \mid \mathbf{v}\right) .
$$



Figure 1: Left: Graph for $p(\mathbf{v}, \mathbf{h})$. Right: Graph for $p(\mathbf{h} \mid \mathbf{v})$
(b) Show that

$$
\begin{equation*}
p\left(h_{i}=1 \mid \mathbf{v}\right)=\frac{1}{1+\exp \left(-b_{i}-\sum_{j} W_{j i} v_{j}\right)} \tag{2}
\end{equation*}
$$

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

Solution. For the conditional $\operatorname{pmf} p\left(h_{i} \mid \mathbf{v}\right)$ 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 $p\left(h_{i} \mid \mathbf{v}\right)$ up to the normalisation constant and then to normalise it afterwards.

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

$$
\begin{align*}
p(\mathbf{h} \mid \mathbf{v}) & =\frac{p(\mathbf{h}, \mathbf{v})}{p(\mathbf{v})}  \tag{S.1}\\
& \propto p(\mathbf{h}, \mathbf{v})  \tag{S.2}\\
& \propto \exp \left(\mathbf{v}^{\top} \mathbf{W h}+\mathbf{a}^{\top} \mathbf{v}+\mathbf{b}^{\top} \mathbf{h}\right)  \tag{S.3}\\
& \propto \exp \left(\mathbf{v}^{\top} \mathbf{W h}+\mathbf{b}^{\top} \mathbf{h}\right)  \tag{S.4}\\
& \propto \exp \left(\sum_{i} \sum_{j} v_{j} W_{j i} h_{i}+\sum_{i} b_{i} h_{i}\right) \tag{S.5}
\end{align*}
$$

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

$$
\begin{equation*}
p\left(h_{i} \mid \mathbf{v}\right) \propto \exp \left(\sum_{j} v_{j} W_{j i} h_{i}+b_{i} h_{i}\right) \tag{S.6}
\end{equation*}
$$

Since $h_{i}$ only takes two values, 0 and 1 , normalisation is here straightforward. Call the unnormalised $\operatorname{pmf} \tilde{p}\left(h_{i} \mid \mathbf{v}\right)$,

$$
\begin{equation*}
\tilde{p}\left(h_{i} \mid \mathbf{v}\right)=\exp \left(\sum_{j} v_{j} W_{j i} h_{i}+b_{i} h_{i}\right) \tag{S.7}
\end{equation*}
$$

We then have

$$
\begin{align*}
p\left(h_{i} \mid \mathbf{v}\right) & =\frac{\tilde{p}\left(h_{i} \mid \mathbf{v}\right)}{\tilde{p}\left(h_{i}=0 \mid \mathbf{v}\right)+\tilde{p}\left(h_{i}=1 \mid \mathbf{v}\right)}  \tag{S.8}\\
& =\frac{\tilde{p}\left(h_{i} \mid \mathbf{v}\right)}{1+\exp \left(\sum_{j} v_{j} W_{j i}+b_{i}\right)}  \tag{S.9}\\
& =\frac{\exp \left(\sum_{j} v_{j} W_{j i} h_{i}+b_{i} h_{i}\right)}{1+\exp \left(\sum_{j} v_{j} W_{j i}+b_{i}\right)} \tag{S.10}
\end{align*}
$$

so that

$$
\begin{align*}
p\left(h_{i}=1 \mid \mathbf{v}\right) & =\frac{\exp \left(\sum_{j} v_{j} W_{j i}+b_{i}\right)}{1+\exp \left(\sum_{j} v_{j} W_{j i}+b_{i}\right)}  \tag{S.11}\\
& =\frac{1}{1+\exp \left(-\sum_{j} v_{j} W_{j i}-b_{i}\right)} \tag{S.12}
\end{align*}
$$

The probability $p(h=0 \mid \mathbf{v})$ equals $1-p\left(h_{i}=1 \mid \mathbf{v}\right)$, which is

$$
\begin{align*}
p\left(h_{i}=0 \mid \mathbf{v}\right) & =\frac{1+\exp \left(\sum_{j} v_{j} W_{j i}+b_{i}\right)}{1+\exp \left(\sum_{j} v_{j} W_{j i}+b_{i}\right)}-\frac{\exp \left(\sum_{j} v_{j} W_{j i}+b_{i}\right)}{1+\exp \left(\sum_{j} v_{j} W_{j i}+b_{i}\right)}  \tag{S.13}\\
& =\frac{1}{1+\exp \left(\sum_{j} W_{j i} v_{j}+b_{i}\right)} \tag{S.14}
\end{align*}
$$

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

$$
p\left(h_{i}=1 \mid \mathbf{v}\right)=\sigma\left(\sum_{j} W_{j i} v_{j}+b_{i}\right) .
$$

(c) Use a symmetry argument to show that

$$
p(\mathbf{v} \mid \mathbf{h})=\prod_{i} p\left(v_{i} \mid \mathbf{h}\right) \quad \text { and } \quad p\left(v_{i}=1 \mid \mathbf{h}\right)=\frac{1}{1+\exp \left(-a_{i}-\sum_{j} W_{i j} h_{j}\right)}
$$

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

$$
\begin{align*}
p(\mathbf{v}, \mathbf{h}) & \propto \exp \left(\mathbf{v}^{\top} \mathbf{W} \mathbf{h}+\mathbf{a}^{\top} \mathbf{v}+\mathbf{b}^{\top} \mathbf{h}\right)  \tag{S.15}\\
& \propto \exp \left(\mathbf{h}^{\top} \mathbf{W}^{\top} \mathbf{v}+\mathbf{b}^{\top} \mathbf{h}+\mathbf{a}^{\top} \mathbf{v}\right) . \tag{S.16}
\end{align*}
$$

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 (2), we thus replace $h_{i}$ with $v_{i}, b_{i}$ with $a_{i}$, and $W_{j i}$ with $W_{i j}$ to obtain $p\left(v_{i}=1 \mid \mathbf{h}\right)$. In terms of the sigmoid function, we have

$$
p\left(v_{i}=1 \mid \mathbf{h}\right)=\sigma\left(\sum_{j} W_{i j} h_{j}+a_{i}\right) .
$$

Note that while $p(\mathbf{v} \mid \mathbf{h})$ 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.

$$
\begin{align*}
p(\mathbf{v}) & =\sum_{\mathbf{h} \in\{0,1\}^{m}} p(\mathbf{v}, \mathbf{h})  \tag{S.17}\\
& =\frac{1}{Z} \sum_{\mathbf{h} \in\{0,1\}^{m}} \exp \left(\mathbf{v}^{\top} \mathbf{W} \mathbf{h}+\mathbf{a}^{\top} \mathbf{v}+\mathbf{b}^{\top} \mathbf{h}\right)  \tag{S.18}\\
& =\frac{1}{Z} \sum_{\mathbf{h} \in\{0,1\}^{m}} \exp \left(\sum_{i j} v_{i} h_{j} W_{i j}+\sum_{i} a_{i} v_{i}+\sum_{j} b_{j} h_{j}\right)  \tag{S.19}\\
& =\frac{1}{Z} \sum_{\mathbf{h} \in\{0,1\}^{m}} \exp \left(\sum_{j=1}^{m} h_{j}\left[\sum_{i} v_{i} W_{i j}+b_{j}\right]+\sum_{i} a_{i} v_{i}\right)  \tag{S.20}\\
& =\frac{1}{Z} \sum_{\mathbf{h} \in\{0,1\}^{m}} \prod_{j=1}^{m} \exp \left(h_{j}\left[\sum_{i} v_{i} W_{i j}+b_{j}\right]\right) \exp \left(\sum_{i} a_{i} v_{i}\right)  \tag{S.21}\\
& =\frac{1}{Z} \exp \left(\sum_{i} a_{i} v_{i}\right) \sum_{\mathbf{h} \in\{0,1\}^{m}} \prod_{j=1}^{m} \exp \left(h_{j}\left[\sum_{i} v_{i} W_{i j}+b_{j}\right]\right)  \tag{S.22}\\
& =\frac{1}{Z} \exp \left(\sum_{i} a_{i} v_{i}\right) \sum_{h_{1}, \ldots, h_{m}} \prod_{j=1}^{m} \exp \left(h_{j}\left[\sum_{i} v_{i} W_{i j}+b_{j}\right]\right) \tag{S.23}
\end{align*}
$$

Importantly, each term in the product only depends on a single $h_{j}$, so that by sequentially applying the distributive law, we have

$$
\begin{align*}
\sum_{h_{1}, \ldots, h_{m}} \prod_{j=1}^{m} \exp \left(h_{j}\left[\sum_{i} v_{i} W_{i j}+b_{j}\right]\right)= & {\left[\sum_{h_{1}, \ldots, h_{m-1}} \prod_{j=1}^{m-1} \exp \left(h_{j}\left[\sum_{i} v_{i} W_{i j}+b_{j}\right]\right)\right] . } \\
& \sum_{h_{m}} \exp \left(h_{m}\left[\sum_{i} v_{i} W_{i m}+b_{m}\right]\right)  \tag{S.24}\\
= & \ldots \\
= & \prod_{j=1}^{m}\left[\sum_{h_{j}} \exp \left(h_{j}\left[\sum_{i} v_{i} W_{i j}+b_{j}\right]\right)\right] \tag{S.25}
\end{align*}
$$

Since $h_{j} \in\{0,1\}$, we obtain

$$
\begin{equation*}
\sum_{h_{j}} \exp \left(h_{j}\left[\sum_{i} v_{i} W_{i j}+b_{j}\right]\right)=1+\exp \left(\sum_{i} v_{i} W_{i j}+b_{j}\right) \tag{S.26}
\end{equation*}
$$

and thus

$$
\begin{equation*}
p(\mathbf{v})=\frac{1}{Z} \exp \left(\sum_{i} a_{i} v_{i}\right) \prod_{j=1}^{m}\left[1+\exp \left(\sum_{i} v_{i} W_{i j}+b_{j}\right)\right] . \tag{S.27}
\end{equation*}
$$

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

$$
\begin{equation*}
Z=\sum_{\mathbf{v} \in\{0,1\}^{n}} \exp \left(\sum_{i} a_{i} v_{i}\right) \prod_{j=1}^{m}\left[1+\exp \left(\sum_{i} v_{i} W_{i j}+b_{j}\right)\right] \tag{S.28}
\end{equation*}
$$

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 $\left(2^{20}=1048576,2^{30}>10^{9}\right)$. Note that different values of $a_{i}, b_{i}, W_{i j}$ yield different values of $Z$. (This is a reason why $Z$ is called the partition function when the $a_{i}, b_{i}, W_{i j}$ are free parameters.)
It is instructive to write $p(\mathbf{v})$ in the log-domain,

$$
\begin{equation*}
\log p(\mathbf{v})=\log Z+\sum_{i=1}^{n} a_{i} v_{i}+\sum_{j=1}^{m} \log \left[1+\exp \left(\sum_{i} v_{i} W_{i j}+b_{j}\right)\right] \tag{S.29}
\end{equation*}
$$

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

$$
\begin{equation*}
f(u)=\log [1+\exp (u)] \tag{S.30}
\end{equation*}
$$

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

$$
\begin{equation*}
\log p(\mathbf{v})=\log Z+\sum_{i=1}^{n} a_{i} v_{i}+\sum_{j=1}^{m} f\left(\sum_{i} v_{i} W_{i j}+b_{j}\right) \tag{S.31}
\end{equation*}
$$

The parameter $b_{j}$ plays the role of a threshold as shown in the figure below. The terms $f\left(\sum_{i} v_{i} W_{i j}+b_{j}\right)$ can be interpreted in terms of feature detection. The sum $\sum_{i} v_{i} W_{i j}$ 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 $f\left(\sum_{i} v_{i} W_{i j}+b_{j}\right)$ a way to measure how much of each feature is present in $\mathbf{v}$.
Further, $\sum_{i} v_{i} W_{i j}+b_{j}$ is also the input to the sigmoid function when computing $p\left(h_{j}=\right.$ $1 \mid \mathbf{v})$. 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 $\sum_{i} v_{i} W_{i j}+b_{j}$ is large for many $j$, i.e. if many features are detected, then $f\left(\sum_{i} v_{i} W_{i j}+b_{j}\right)$ will be non-zero for many $j$, and $\log p(\mathbf{v})$ will be large.


