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

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

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

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 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

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

Figure 1: Left: Graph for $p(\mathbf{v}, \mathbf{h})$. Right: Graph for $p(\mathbf{h}|\mathbf{v})$

(b) Show that

$$p(h_i = 1|\mathbf{v}) = \frac{1}{1 + \exp\left(-b_i - \sum_j W_{ji} v_j\right)} \tag{2}$$

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

$$p(\mathbf{h}|\mathbf{v}) = \frac{p(\mathbf{h},\mathbf{v})}{p(\mathbf{v})} \tag{S.1}$$

$$\propto p(\mathbf{h},\mathbf{v}) \tag{S.2}$$

$$\propto \exp\left(\mathbf{v}^\top \mathbf{W}\mathbf{h} + \mathbf{a}^\top \mathbf{v} + \mathbf{b}^\top \mathbf{h}\right) \tag{S.3}$$

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

$$\propto \exp\left(\sum_i \sum_j v_j W_{ji} h_i + \sum_i b_i h_i\right) \tag{S.5}$$

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

$$p(h_i|\mathbf{v}) \propto \exp\left(\sum_j v_j W_{ji} h_i + b_i h_i\right) \tag{S.6}$$

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

$$\tilde{p}(h_i|\mathbf{v}) = \exp\left(\sum_j v_j W_{ji} h_i + b_i h_i\right). \tag{S.7}$$

We then have

$$p(h_i|\mathbf{v}) = \frac{\tilde{p}(h_i|\mathbf{v})}{\tilde{p}(h_i=0|\mathbf{v}) + \tilde{p}(h_i=1|\mathbf{v})} \tag{S.8}$$

$$= \frac{\tilde{p}(h_i|\mathbf{v})}{1 + \exp\left(\sum_j v_j W_{ji} + b_i\right)} \tag{S.9}$$

$$= \frac{\exp\left(\sum_j v_j W_{ji} h_i + b_i h_i\right)}{1 + \exp\left(\sum_j v_j W_{ji} + b_i\right)}, \tag{S.10}$$

so that

$$p(h_i=1|\mathbf{v}) = \frac{\exp\left(\sum_j v_j W_{ji} + b_i\right)}{1 + \exp\left(\sum_j v_j W_{ji} + b_i\right)} \tag{S.11}$$

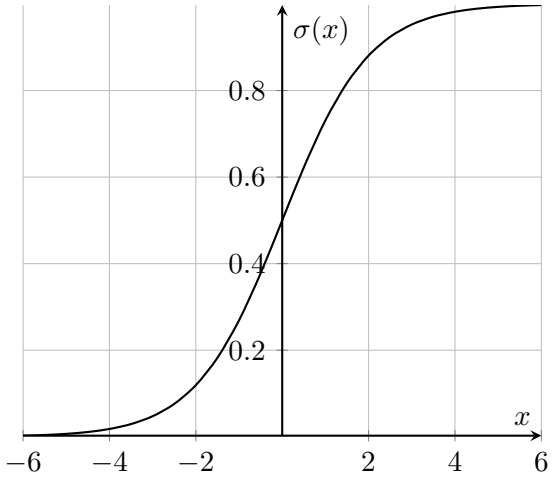
$$= \frac{1}{1 + \exp\left(-\sum_j v_j W_{ji} - b_i\right)}. \tag{S.12}$$

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

$$p(h_i=0|\mathbf{v}) = \frac{1 + \exp\left(\sum_j v_j W_{ji} + b_i\right)}{1 + \exp\left(\sum_j v_j W_{ji} + b_i\right)} - \frac{\exp\left(\sum_j v_j W_{ji} + b_i\right)}{1 + \exp\left(\sum_j v_j W_{ji} + b_i\right)} \tag{S.13}$$

$$= \frac{1}{1 + \exp\left(\sum_j W_{ji} v_j + b_i\right)} \tag{S.14}$$

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

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

$$\propto \exp\left(\mathbf{h}^\top \mathbf{W}^\top \mathbf{v} + \mathbf{b}^\top \mathbf{h} + \mathbf{a}^\top \mathbf{v}\right). \tag{S.16}$$

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_{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.

$$p(\mathbf{v}) = \sum_{\mathbf{h}\in\{0,1\}^m} p(\mathbf{v},\mathbf{h}) \tag{S.17}$$

$$= \frac{1}{Z}\sum_{\mathbf{h}\in\{0,1\}^m} \exp\left(\mathbf{v}^\top\mathbf{W}\mathbf{h} + \mathbf{a}^\top\mathbf{v} + \mathbf{b}^\top\mathbf{h}\right) \tag{S.18}$$

$$= \frac{1}{Z}\sum_{\mathbf{h}\in\{0,1\}^m} \exp\left(\sum_{ij} v_i h_j W_{ij} + \sum_i a_i v_i + \sum_j b_j h_j\right) \tag{S.19}$$

$$= \frac{1}{Z}\sum_{\mathbf{h}\in\{0,1\}^m} \exp\left(\sum_{j=1}^{m} h_j\left[\sum_i v_i W_{ij} + b_j\right] + \sum_i a_i v_i\right) \tag{S.20}$$

$$= \frac{1}{Z}\sum_{\mathbf{h}\in\{0,1\}^m} \prod_{j=1}^{m} \exp\left(h_j\left[\sum_i v_i W_{ij} + b_j\right]\right) \exp\left(\sum_i a_i v_i\right) \tag{S.21}$$

$$= \frac{1}{Z}\exp\left(\sum_i a_i v_i\right) \sum_{\mathbf{h}\in\{0,1\}^m} \prod_{j=1}^{m} \exp\left(h_j\left[\sum_i v_i W_{ij} + b_j\right]\right) \tag{S.22}$$

$$= \frac{1}{Z}\exp\left(\sum_i a_i v_i\right) \sum_{h_1,\ldots,h_m} \prod_{j=1}^{m} \exp\left(h_j\left[\sum_i v_i W_{ij} + b_j\right]\right) \tag{S.23}$$

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

$$\sum_{h_1,\ldots,h_m} \prod_{j=1}^{m} \exp\left(h_j\left[\sum_i v_i W_{ij} + b_j\right]\right) = \left[\sum_{h_1,\ldots,h_{m-1}} \prod_{j=1}^{m-1} \exp\left(h_j\left[\sum_i v_i W_{ij} + b_j\right]\right)\right] \cdot \sum_{h_m} \exp\left(h_m\left[\sum_i v_i W_{im} + b_m\right]\right) \tag{S.24}$$

$$= \ldots$$

$$= \prod_{j=1}^{m}\left[\sum_{h_j} \exp\left(h_j\left[\sum_i v_i W_{ij} + b_j\right]\right)\right] \tag{S.25}$$

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

$$\sum_{h_j} \exp\left(h_j\left[\sum_i v_i W_{ij} + b_j\right]\right) = 1 + \exp\left(\sum_i v_i W_{ij} + b_j\right) \tag{S.26}$$

and thus

$$p(\mathbf{v}) = \frac{1}{Z}\exp\left(\sum_i a_i v_i\right) \prod_{j=1}^{m}\left[1 + \exp\left(\sum_i v_i W_{ij} + b_j\right)\right]. \tag{S.27}$$

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

$$Z = \sum_{\mathbf{v}\in\{0,1\}^n} \exp\left(\sum_i a_i v_i\right) \prod_{j=1}^{m} \left[1 + \exp\left(\sum_i v_i W_{ij} + b_j\right)\right] \tag{S.28}$$

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.)

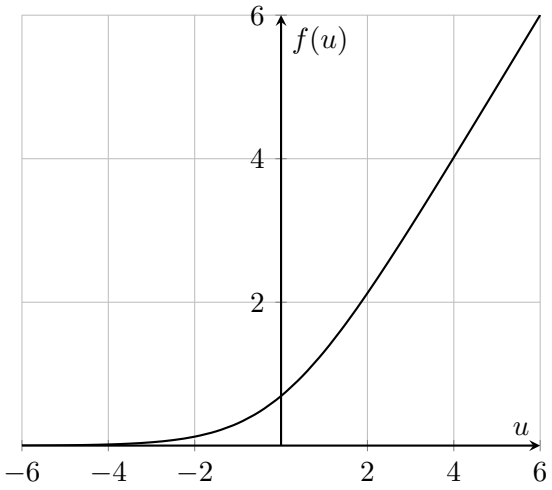
It is instructive to write $p(\mathbf{v})$ in the log-domain,

$$\log p(\mathbf{v}) = \log Z + \sum_{i=1}^{n} a_i v_i + \sum_{j=1}^{m} \log\left[1 + \exp\left(\sum_i v_i W_{ij} + b_j\right)\right], \tag{S.29}$$

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

$$f(u) = \log\left[1 + \exp(u)\right], \tag{S.30}$$

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)



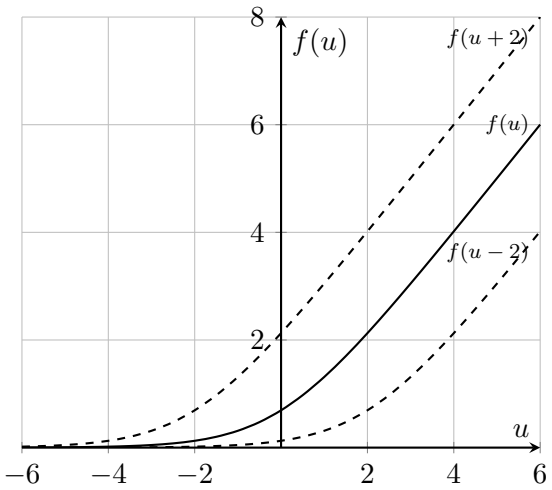
With the softplus function $f(u)$, we can write $\log p(\mathbf{v})$ as

$$\log p(\mathbf{v}) = \log Z + \sum_{i=1}^{n} a_i v_i + \sum_{j=1}^{m} f\left(\sum_i v_i W_{ij} + b_j\right). \tag{S.31}$$

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.


8
6
4
2
$f(u)$
$f(u+2)$
$f(u)$
$f(u-2)$
u
-6
-4
-2
2
4
6