Exercise 1: In the context of RBMs, write all the steps to prove that

$$P(h|v) = \prod_{j=1}^{m_h} \sigma\left((2h-1) \odot (c + W^T v)\right)_j .$$

SOLUTION: Remember that we are considering a (shallow) RBM with m_h BINARY HIDDEN units, which we refer to collectively with the vector h, and m_v BINARY VISIBLE units, which we refer to collectively with the vector v.

Remember also that $P(v,h) = \frac{1}{Z} \exp\{b^T v + c^T h + v^T W h\}$.

$$P(h|v) = \frac{P(v,h)}{P(v)} \qquad [\text{PRODUCT RULE: } P(v,h) = P(h|v) \cdot P(v)]$$

$$= \frac{1}{P(v)} \cdot \frac{1}{Z} \exp\{b^T v + c^T h + v^T W h\}$$

$$= \frac{1}{P(v)} \cdot \frac{1}{Z} \exp\{b^T v\} \cdot \exp\{c^T h + v^T W h\} \qquad [\exp(a+b) = \exp(a) \cdot \exp(b)]$$

$$\begin{bmatrix} v \text{ is CONSTANT} \\ \text{WRT } P(h|v) \end{bmatrix} = \frac{1}{Z'} \cdot \exp\{c^T h + v^T W h\} \qquad [\text{with } Z' = P(v) \cdot Z \cdot \exp\{-b^T v\}]$$

$$= \frac{1}{Z'} \cdot \exp\left\{ \sum_{j=1}^{m_h} c_j h_j + \sum_{j=1}^{m_h} v^T \underbrace{W_{:,j}}_{j\text{-th COLUMN of } W} h_j \right\}$$

$$= \frac{1}{\underset{\text{CONSTANT}}{Z'}} \cdot \underbrace{\prod_{j=1}^{m_h} \exp\{c_j h_j + v^T W_{:,j} h_j\}}_{\text{PRODUCT OF UNNORMALIZED DISTRIBUTIONS OVER THE INDIVIDUAL ELEMENTS } h_j:\ \tilde{P}(h_j|v)} \qquad \left[\exp\left\{\sum_j [\dots]\right\} = \prod_j \exp\{[\dots]\}\right]$$

$\frac{1}{Z'}$ is the NORMALIZATION FACTOR

NOW, WE NORMALIZE the DISTRIBUTIONS over the individual BINARY h_j:

$$P(h_j = 1|v) = \frac{\tilde{P}(h_j = 1|v)}{\boxed{\tilde{P}(h_j = 0|v) + \tilde{P}(h_j = 1|v)}} \leftarrow \text{TO NORMALIZE} \qquad \tilde{P} = \text{UNNORMALIZED DIST.}$$

$$= \frac{\exp\{c_j + v^T W_{:,j}\}}{\exp\{0\} + \exp\{c_j + v^T W_{:,j}\}}$$

$$= \sigma(c_j + v^T W_{:,j}), \quad \text{because } \sigma(x) = \frac{\exp(x)}{1 + \exp(x)}$$

We also have to compute $P(h_j=0|v)$. We can exploit that

$$P(h_j=0|v) = 1 - P(h_j=1|v) = 1 - \sigma(c_j + v^T W_{:,j}) \underset{\uparrow}{=} \sigma(-c_j - v^T W_{:,j})$$

because $1-\sigma(x) = \sigma(-x)$

Finally: $$P(h|v) = \prod_{j=1}^{m_h} P(h_j|v) = \prod_{j=1}^{m_h} \sigma\big(\underbrace{(2h-1)\odot(c+W^T v)}_{\text{VECTOR FORM}}\big)_j$$

This is a COMPACT FORM that takes into account all possible values of each h_j (0, 1). In fact:

- If $h=0 \Rightarrow 2h-1=-1 \Rightarrow \sigma(-c-W^T v)_j = \sigma(-c_j - v^T W_{:,j})$
- If $h=1 \Rightarrow 2h-1=1 \Rightarrow \sigma(c+W^T v)_j = \sigma(c_j + v^T W_{:,j})$

Note 1: $(2h-1)\odot(c+W^T v) \in \mathbb{R}^{m_h}$, indeed:

$$h \in \mathbb{R}^{m_h},\ c \in \mathbb{R}^{m_h},\ W \in \mathbb{R}^{m_v \times m_h} \Rightarrow W^T \in \mathbb{R}^{m_h \times m_v} \Rightarrow W^T \underset{\mathbb{R}^{m_v}}{\underset{\cap}{v}} \in \mathbb{R}^{m_h}$$

Note 2: why do we go from $v^T W_{:,j}$ to $(W^T v)_j$?

Because $\underset{\mathbb{R}^{1\times m_v}}{\underset{\cap}{v^T}} \underset{\mathbb{R}^{m_v \times m_h}}{\underset{\cap}{W}} \in \mathbb{R}^{1\times m_h} \Rightarrow$ is a ROW VECTOR.

When we write things in vector form, we consider COLUMN VECTOR.
To obtain a column vector from $v^T W$, we can TRANSPOSE it:

$$(v^T W)^T \underset{\substack{\uparrow \\ (A\cdot B)^T = B^T\cdot A^T}}{=} W^T (v^T)^T \underset{\substack{\uparrow \\ (A^T)^T = A}}{=} W^T v \in \mathbb{R}^{m_h}$$

$\underset{\mathbb{R}^{m_h \times m_v}}{\underset{\cap}{W^T}} \cdot \underset{\mathbb{R}^{m_v}}{\underset{\cap}{v}} \rightarrow$ the product is a COLUMN vector of dim. m_h.

Exercise 2: In the context of RBMs, write all the steps to prove that:

$$P(v|h) = \prod_{i=1}^{m_v} \sigma\left((2v-1) \odot (b + Wh)\right)_i$$

SOLUTION:

$$P(v|h) = \frac{P(v,h)}{P(h)} \qquad \text{[PRODUCT RULE]}$$

$$= \frac{1}{P(h)} \cdot \frac{1}{Z} \cdot \exp\{b^T v + c^T h + v^T W h\}$$

[h is constant wrt P(v|h)]

$$= \frac{1}{P(h)} \cdot \frac{1}{Z} \cdot \exp\{c^T h\} \cdot \exp\{b^T v + v^T W h\}$$

$$= \frac{1}{Z'} \cdot \exp\{b^T v + v^T W h\} \qquad \left(\text{with } Z' = P(h) \cdot Z \cdot \exp\{c^T h\}\right)$$

$$= \frac{1}{Z'} \cdot \exp\left\{\sum_{i=1}^{m_v} b_i v_i + \sum_{i=1}^{m_v} v_i \underbrace{W_{i,:}}_{i\text{-th ROW of } W} h\right\}$$

$$= \frac{1}{Z'} \cdot \prod_{i=1}^{m_v} \underbrace{\exp\{b_i v_i + v_i W_{i,:} h\}}_{\text{UNNORMALIZED PROBABILITIES } \tilde{P}(v_i|h)}$$

Again, normalize the distributions over the individual BINARY v_i:

$$P(v_i = 1|h) = \frac{\tilde{P}(v_i = 1|h)}{\underbrace{\tilde{P}(v_i = 0|h) + \tilde{P}(v_i = 1|h)}_{\text{TO NORMALIZE}}} \qquad \tilde{P} = \text{UNNORMALIZED DIST.}$$

$$= \frac{\exp\{b_i + W_{i,:} h\}}{\exp\{0\} + \exp\{b_i + W_{i,:} h\}}$$

$$= \sigma(b_i + W_{i,:} h)$$

Compute also $P(v_i = 0|h)$ as:

$$P(v_i = 0|h) = 1 - P(v_i = 1|h) = 1 - \sigma(b_i + W_{i,:} h) = \sigma(-b_i - W_{i,:} h)$$

$$\Rightarrow P(v|h) = \prod_{i=1}^{m_v} P(v_i|h) = \prod_{i=1}^{m_v} \sigma\Big(\underbrace{(2v-1) \odot (b + Wh)}_{\text{VECTOR FORM that encodes the two possible values for } v_i\text{: } 0,\ 1.}\Big)_i$$

Exercise: Given stochastic variables X_1, X_2, X_3, X_4 and the fact that

- $X_2 \perp\!\!\!\perp X_3 \mid X_4$
- $X_1 \perp\!\!\!\perp X_4 \mid X_2$

the joint prob. distribution can be factorized as:

a. $P(X_2)\,P(X_4|X_2)\,P(X_1|X_2)\,P(X_3|X_1,X_2,X_4)$

b. $P(X_1)\,P(X_2)\,P(X_3)\,P(X_4|X_3,X_2)$

c. $P(X_3)\,P(X_2|X_1)\,P(X_3|X_1)\,P(X_4|X_2)$

d. $P(X_1)\,P(X_2|X_1)\,P(X_3|X_2)\,P(X_4|X_3)$

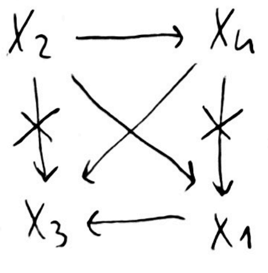
e. $P(X_3)\,P(X_2|X_1)\,P(X_3|X_1)\,P(X_4|X_2)$

Solution:

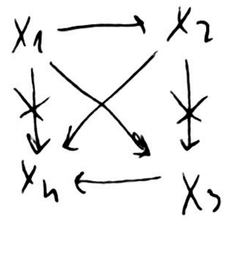
1) Proceed by exclusion.

c) and e) are the same → NOT CORRECT ANSWERS.

b) NOT CORRECT because there are not sufficient conditional independencies to obtain all these marginals.

a) 

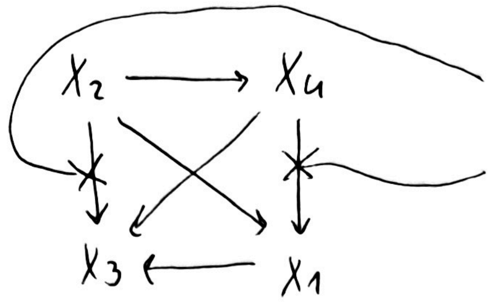
$$\underbrace{P(X_2)\,P(X_4|X_2)\,P(X_1|X_2)\,P(X_3|X_1,X_4)}_{\text{the closest solution would be a)}}$$

d) 

$P(X_1)\,P(X_2|X_1)\,P(X_3|X_1)\,P(X_4|X_2,X_3)$

not close to the answer (d)

Let's start again with the complete graph:



$X_2 \perp\!\!\!\perp X_3 \mid X_4$

$X_1 \perp\!\!\!\perp X_4 \mid X_2$

So, our factorization would be:

$$(1) \quad P(X_1, X_2, X_3, X_4) = \underbrace{P(X_2) \cdot P(X_4 \mid X_2)\, P(X_1 \mid X_2)}_{\parallel \atop P(X_1, X_2, X_4)}\, P(X_3 \mid X_1, X_4)$$

Indeed:

$$P(X_1, X_2, X_4) \overset{\text{PROD RULE}}{=} P(X_2) \cdot P(X_1, X_4 \mid X_2)$$

$$(X_1 \perp\!\!\!\perp X_4 \mid X_2) = P(X_2) \cdot P(X_1 \mid X_2) \cdot P(X_4 \mid X_2) \quad \blacksquare$$

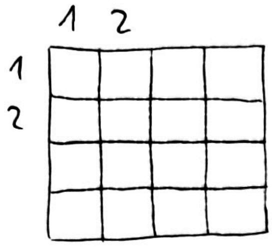
In particular:

$$(2) \quad P(X_1, X_2, X_3, X_4) = P(X_1, X_2, X_4) \cdot P(X_3 \mid X_1, X_2, X_4) \qquad \text{prod. rule}$$

From ① = ② we obtain $P(X_3 \mid X_1, X_4) = P(X_3 \mid X_1, X_2, X_4)$

Exercise: Consider an UNSHARED CNN layer with 1 filter of size 3X3, a stride of 1, input image 4X4 and no padding. How many PARAMETERS are we required to train such a layer? Do not consider the bias term.

Solution.



$3 \times 3 \times 4 = 36.$

Note: Unshared CNNs are for STRUCTURED, LOCATION-DEPENDANT DATA.

- 10 FILTERS: $36 \times 10 = 360.$

- 1 FILTER + 2 INPUT CHANNELS: $36 \times 2 = 72.$

- INCLUDING BIAS: $36 + 4 = 40.$ (one bias term $\forall$ filter)

- PADDING: $3 \times 3 \times 4 \times 4 = 144.$

- PADDING + BIAS: $144 + 16 = 160.$

- STRIDE = 2: $3 \times 3 = 9.$

Hint: the number of PARAMETERS in an UNSHARED CNN is equal to the number of MULTIPLICATIONS in a SHARED CNN, given same filters, input...