

# ADVANCED MACHINE LEARNING

## EXERCISE #2 SOLUTION

Uri Avron [uriavron@gmail.com] [308046994]  
Jonathan Somer [jonathan.somer@gmail.com] [307923383]  
Matan Harel [matan.harel.mh@gmail.com] [302695721]

April 15, 2018

### Question 2

**Show that the tree-width of a 2D graph of size  $(m,m)$  is at most  $m$**

Recall tree-width is defined as:

“The minimum size of the maximum clique [in the induced graph], over all elimination orders, minus one”.

Thus, our strategy will be to show an elimination order which has a maximal clique of size  $m$ . The minimum over all elimination orders cannot exceed this value so the tree-width of the grid will be at most  $m$  (well,  $m - 1$ , but “at most  $m$ ” holds too). We shall show that eliminating columns of the grid, top to bottom, sequentially results in a maximal clique of size  $m$  over the entire elimination.

Denote the variables of column  $i$  as  $\{x_1^i, \dots, x_m^i\}$ .

**Removing the first column:**

After removing  $x_1^1$  we end up with:

$$\tau(x_2^1, x_1^2) \prod_{i,j \in E \setminus \{\text{edges involving } x_1^1\}} \phi(x_i, x_j)$$

After removal of  $x_k^1$  for  $k < m$ :

$$\tau(x_{k+1}^1, x_k^2, \dots, x_1^2) \prod_{i,j \in E \setminus \{\text{edges involving } \{x_1^1, \dots, x_k^1\}\}} \phi(x_i, x_j)$$

Removing  $x_{m-1}^1$ , we get a factor of size  $m$ , and will thus have a clique of size  $m$ . We will show that during the rest of the process no factor larger than  $m$

will appear. Continuing:  
After removal of  $x_m^1$  we have:

$$\tau(x_m^2, \dots, x_1^2) \prod_{i,j \in E \setminus \{\text{edges involving } \{x_1^1, \dots, x_m^1\}\}} \phi(x_i, x_j) \quad (1)$$

**Removing columns 2 to m-1:**

second column, removing  $x_1^2$  we arrive at:

$$\tau(x_m^2, \dots, x_1^3) \prod_{i,j \in E \setminus \{\text{edges involving } \{x_1^1, \dots, x_m^1, x_1^2\}\}} \phi(x_i, x_j)$$

Removing  $x_k^2$  for  $k < m$  :

$$\tau(x_m^2, \dots, x_{k+1}^2, x_k^3, \dots, x_1^3) \prod_{i,j \in E \setminus \{\text{edges involving } \{x_1^1, \dots, x_m^1, x_1^2, \dots, x_k^2\}\}} \phi(x_i, x_j)$$

Removing  $x_m^2$  :

$$\tau(x_m^3, \dots, x_1^3) \prod_{i,j \in E \setminus \{\text{edges involving } \{x_1^1, \dots, x_m^1, x_1^2, \dots, x_k^2\}\}} \phi(x_i, x_j)$$

Note: at this point we arrive at a structure similar to (1). We continue this way, repeating the same steps as in removing the second column until we arrive at the last column.

**Removing the last column:**

At this point we have:

$$\tau(x_m^m, \dots, x_1^m)$$

The last column is removed one by one as the single remaining factor is replaced by one with one variable less at each stage.

In total we have seen a maximal factor of size  $m$ . As we know this corresponds with a maximal clique size of size  $m$ . Thus, the tree-width cannot exceed  $m$ .

### Question 3

**“Consider the sum-product message update on a tree graph. But, consider the case where all messages are updated simultaneously. Namely:**

$$m_{ij}^{t+1}(x_j) = \sum_{x_i} \phi(x_i, x_j) \prod_{k \in N(i) \setminus \{j\}} m_{ki}^t(x_i)$$

**Show that this converges to the true marginals at iteration  $t = n$ .**

Initialization:

$$\forall ij \in E : m_{ij}^0 = \begin{cases} \sum_{x_i} \phi(x_i, x_j) & \text{if } i \text{ is a leaf} \\ 0 & \text{else} \end{cases}$$

Recall our definition for a message between  $i$  and  $j$ :

$$m_{ij}(x_j) = \sum_{x_i} \phi(x_i, x_j) \prod_{k \in N(i) \setminus \{j\}} m_{ki}(x_i) \quad (2)$$

**Claim:**

At iteration  $t$ :  $m_{ij}^t$  holds the correct message value, function (2) we defined above, for all nodes whose longest path to any descendent is at most  $t$ . For all other nodes  $m_{ij}^t = 0$  (0 as a constant function).

[ Note: the notion of “children of a node” is on a per-message basis. For message  $m_{ij}^t$  from node  $i$  to node  $j$ , we “treat”  $j$  as the root and the descendents of node  $i$  are defined as such. ]

**Proof of claim - by induction:**

**Base case  $t = 0$ .** Leaves are the only nodes “whose longest path to any child is at most 0”. Because leaves have only one neighbour - their parent, they do not receive any messages and the value we have computed for leaves upon initialization is the correct value of their message. For all non-leaf nodes, initialization was to 0 and the claim holds.

**Assume for  $k - 1$ .** We now calculate  $m_{ij}^k$  as defined. Let  $x_i$  be some node whose longest path to any descendent is at most  $k$ , the longest path from any of  $x_i$ ’s children to any of their descendents is at most  $k - 1$ . By the inductive hypothesis their messages have been calculated correctly. Thus the update has placed the correct value in  $m_{ij}^k$ . For all other nodes, some child of theirs holds 0 and the product will be have a 0 factor for any value of  $x_i$ .  $\square$

**Main proof:**

Let  $p(x_i)$  be some marginal we wish to compute. From the claim above and the fact that for any tree of size  $n$  the distance from root to any of its descendents is at most  $n$ , it follows that all messages have been computed correctly. Thus, the process has converged to the true marginals and  $p(x_i)$  is given by:

$$p(x_i) = \prod_{k \in N(i)} m_{ki}^n(x_i)$$

$\square$

## Question 4

Define:

$$\begin{aligned}\mu_i(x_i) &\propto \prod_{k \in N(i)} m_{ki}(x_i) \\ \mu_{ij}(x_i, x_j) &\propto \phi(x_i, x_j) \prod_{k \in N(i) \setminus j} m_{ki}(x_i) \prod_{k \in N(j) \setminus i} m_{kj}(x_j)\end{aligned}$$

(a)

We should show that  $p(x_i, x_j) = \mu_{ij}(x_i, x_j)$  when  $G$  is a tree.

The marginal of  $p(x_i, x_j)$  of an edge  $(i, j) \in E(G)$  is:

$$p(x_i, x_j) = \sum_{x \setminus \{x_i, x_j\}} p(x) = \frac{1}{Z} \sum_{x \setminus \{x_i, x_j\}} \prod_{(k,l) \in E} \phi_{kl}(x_k, x_l)$$

Since  $G$  is a tree, if we remove the edge  $(x_i, x_j)$  two sub trees will be created with no path between  $x_i$  and  $x_j$ .

We denote the subtree with  $x_i$  as  $G_i$  and the subtree with  $x_j$  as  $G_j$ . so the above equation can be written as:

$$\begin{aligned}p(x_i, x_j) &= \frac{1}{Z} \sum_{x \setminus \{x_i, x_j\}} \phi_{i,j}(x_i, x_j) \prod_{(k,l) \in E(G_i)} \phi_{k,l}(x_k, x_l) \prod_{(k,l) \in E(G_j)} \phi_{k,l}(x_k, x_l) \\ &= \frac{1}{Z} \times \phi_{i,j}(x_i, x_j) \times \sum_{x \setminus \{x_i, x_j\}} \left( \prod_{(k,l) \in E(G_i)} \phi_{k,l}(x_k, x_l) \times \prod_{(k,l) \in E(G_j)} \phi_{k,l}(x_k, x_l) \right) \\ &= \frac{1}{Z} \times \phi_{i,j}(x_i, x_j) \times \sum_{x \setminus \{x_i, x_j\}} \prod_{(k,l) \in E(G_i)} \phi_{k,l}(x_k, x_l) \times \sum_{x \setminus \{x_i, x_j\}} \prod_{(k,l) \in E(G_j)} \phi_{k,l}(x_k, x_l)\end{aligned}$$

Lets divide each subtree  $G_i$  and  $G_j$  once more to its neighbors sub-trees

$$\begin{aligned}p(x_i, x_j) &= \frac{1}{Z} \times \phi_{i,j}(x_i, x_j) \\ &\quad \times \prod_{d \in N(i) \setminus \{j\}} \sum_{s \in N(d) \setminus \{i\}} \prod_{(k,l) \in E(G_s)} \phi_{k,l}(x_k, x_l) \\ &\quad \times \prod_{d \in N(j) \setminus \{i\}} \sum_{s \in N(d) \setminus \{j\}} \prod_{(k,l) \in E(G_s)} \phi_{k,l}(x_k, x_l)\end{aligned}$$

Since  $m_{di}(x_i) = \sum_{s \in N(d) \setminus \{i\}} \prod_{(k,l) \in E(G_s)} \phi_{k,l}(x_k, x_l)$  we get from the definition of

$\mu_{i,j}(x_i, x_j)$ :

$$p(x_i, x_j) = \frac{1}{Z} \phi_{i,j}(x_i, x_j) \times \prod_{d \in N(i) \setminus \{j\}} m_{di}(x_i) \times \prod_{d \in N(j) \setminus \{i\}} m_{dj}(x_j) = \mu_{i,j}(x_i, x_j)$$

□

(b)

On Markov net  $p(x_{[n]}) \propto \prod_{(i,j) \in E} \phi_{i,j}(x_i, x_j)$  so we will show that

$$\prod_{(i,j) \in E} \phi_{i,j}(x_i, x_j) = \prod_i \mu_i(x_i) \times \prod_{(i,j) \in E} \frac{\mu_{i,j}(x_i, x_j)}{\mu_i(x_i) \mu_j(x_j)}$$

Lets develop the right expression:

$$\begin{aligned} \prod_{(i,j) \in E} \frac{\mu_{i,j}(x_i, x_j)}{\mu_i(x_i) \mu_j(x_j)} &= \prod_{(i,j) \in E} \frac{\phi_{i,j}(x_i, x_j) \times \prod_{d \in N(i) \setminus \{j\}} m_{di}(x_i) \times \prod_{d \in N(j) \setminus \{i\}} m_{dj}(x_j)}{\mu_i(x_i) \mu_j(x_j)} \\ &= \prod_{(i,j) \in E} \frac{\phi_{i,j}(x_i, x_j) \times \prod_{d \in N(i) \setminus \{j\}} m_{di}(x_i) \times \prod_{d \in N(j) \setminus \{i\}} m_{dj}(x_j)}{\prod_{d \in N(i)} m_{di}(x_i) \times \prod_{d \in N(j)} m_{dj}(x_j)} \\ &= \prod_{(i,j) \in E} \frac{\phi_{i,j}(x_i, x_j)}{m_{ij}(x_j) \times m_{ji}(x_i)} \\ &= \frac{\prod_{(i,j) \in E} \phi_{i,j}(x_i, x_j)}{\prod_i \prod_{j \in N(i)} m_{ji}(x_i)} \\ &= \frac{\prod_{(i,j) \in E} \phi_{i,j}(x_i, x_j)}{\prod_i \mu_i(x_i)} \end{aligned}$$

Place the new form in the first equation to get the result we are looking for:

$$\prod_i \mu_i(x_i) \times \prod_{(i,j) \in E} \frac{\mu_{i,j}(x_i, x_j)}{\mu_i(x_i) \mu_j(x_j)} = \prod_i \mu_i(x_i) \times \frac{\prod_{(i,j) \in E} \phi_{i,j}(x_i, x_j)}{\prod_i \mu_i(x_i)} = \prod_{(i,j) \in E} \phi_{i,j}(x_i, x_j)$$

Therefore  $p(x_1 \dots x_n) \propto \prod_i \mu_i(x_i) \times \prod_{(i,j) \in E} \frac{\mu_{i,j}(x_i, x_j)}{\mu_i(x_i) \mu_j(x_j)}$

□

(c)

We will prove that at a fixed point of LBP we get

$$\frac{\sum_{x_j} \mu_{i,j}(x_i, x_j)}{\mu_i(x_i)} = 1$$

Lets start with the upper expression, lets say that the normalization factor of  $\mu_{ij}$  is  $D_1$ . so we can rewrite the expression:

$$\begin{aligned} \sum_{x_j} \mu_{ij}(x_i, x_j) &= D_1 \times \sum_{x_j} \phi_{i,j}(x_i, x_j) \times \prod_{d \in N(i) \setminus \{j\}} m_{di}(x_i) \times \prod_{d \in N(j) \setminus \{i\}} m_{dj}(x_j) \\ &= D_1 \times \prod_{d \in N(i) \setminus \{j\}} m_{di}(x_i) \times \left( \sum_{x_j} \phi_{i,j}(x_i, x_j) \prod_{d \in N(j) \setminus \{i\}} m_{dj}(x_j) \right) \end{aligned}$$

On the other hand, lets open the lower expression according to the definition of  $\mu_i(x_i)$  and assume the normalization factor is  $D_2$ :

$$\mu_i(x_i) = D_2 \times \prod_{d \in N(i)} m_{di}(x_i)$$

Combine it together we get:

$$\begin{aligned} \frac{\sum_{x_j} \mu_{i,j}(x_i, x_j)}{\mu_i(x_i)} &= \frac{D_1 \times \prod_{d \in N(j) \setminus \{i\}} m_{dj}(x_j) \times \left( \sum_{x_j} \phi_{i,j}(x_i, x_j) \prod_{d \in N(j) \setminus \{i\}} m_{dj}(x_j) \right)}{D_2 \times \prod_{d \in N(i)} m_{di}(x_i)} \\ &= \frac{D_1}{D_2} \times \frac{\sum_{x_j} \phi_{i,j}(x_i, x_j) \prod_{d \in N(j) \setminus \{i\}} m_{dj}(x_j)}{m_{ji}(x_i)} \end{aligned}$$

Now lets open the  $m_{ji}(x_i)$  expression, and suppose the normalization factor is  $D_3$

$$\frac{\sum_{x_j} \mu_{i,j}(x_i, x_j)}{\mu_i(x_i)} = \frac{D_1}{D_2} \times \frac{\sum_{x_j} \phi_{i,j}(x_i, x_j) \prod_{d \in N(j) \setminus \{i\}} m_{dj}(x_j)}{D_3 \times \sum_{x_j} \phi_{i,j}(x_i, x_j) \times \prod_{d \in N(j) \setminus \{i\}} m_{dj}(x_j)} = \frac{D_1}{D_2 \times D_3}$$

We've got that the value of this expression depends just on the normalization factors, but the normalization factors sum to 1 over all values. therefore

$$\frac{\sum_{x_j} \mu_{i,j}(x_i, x_j)}{\mu_i(x_i)} = 1 \Rightarrow \sum_{x_j} \mu_{i,j}(x_i, x_j) = \mu_i(x_i)$$

□