Prof. Ryan Cotterell

Simon Wachter: Assignment 2

siwachte@ethz.ch, 19-920-198

04/12/2022 - 11:41h

Question 1:

a) Prove that the expectation semiring satisfies the semiring axioms:

- $(\mathbb{R} \times \mathbb{R}, \oplus, \mathbf{0})$ must be a commutative monoid with identity element $\mathbf{0}$:

$$(\langle x, y \rangle \oplus \langle x', y' \rangle) \oplus \langle x'', y'' \rangle = \langle x + x', y + y' \rangle \oplus \langle x'', y'' \rangle \tag{1}$$
$$= \langle x + x' + x'', y + y' + y'' \rangle \tag{2}$$
$$= \langle x, y \rangle \oplus \langle x' + x'', y' + y'' \rangle \tag{3}$$
$$= \langle x, y \rangle \oplus (\langle x', y' \rangle \oplus \langle x'', y'' \rangle) \tag{4}$$

$$\mathbf{0} + \langle x, y \rangle = \langle 0, 0 \rangle \oplus \langle x, y \rangle \tag{5}$$
$$= \langle 0 + x, 0 + y \rangle \tag{6}$$
$$= \langle x, y \rangle \tag{7}$$
$$= \langle x + 0, y + 0 \rangle \tag{8}$$
$$= \langle x, y \rangle + \mathbf{0} \tag{9}$$

$$\langle x, y \rangle + \langle x', y' \rangle = \langle x + x', y + y' \rangle \tag{10}$$
$$= \langle x' + x, y' + y \rangle \tag{11}$$
$$= \langle x', y' \rangle + \langle x, y \rangle \tag{12}$$

- $(\mathbb{R} \times \mathbb{R}, \otimes, \mathbf{1})$ must be a monoid with identity element $\mathbf{1}$:

$$(\langle x, y \rangle \otimes \langle x', y' \rangle) \otimes \langle x'', y'' \rangle = \langle x \cdot x', x \cdot y' + y \cdot x' \rangle \otimes \langle x'', y'' \rangle \tag{13}$$
$$= \langle x \cdot x' \cdot x'', x \cdot x' \cdot y'' + (x \cdot y' + y \cdot x') \cdot x'' \rangle \tag{14}$$
$$= \langle x \cdot x' \cdot x'', x \cdot x' \cdot y'' + x \cdot y' \cdot x'' + y \cdot x' \cdot x'' \rangle \tag{15}$$
$$= \langle x, y \rangle \otimes \langle x' \cdot x'', x' \cdot y'' + y' \cdot x'' \rangle \tag{16}$$
$$= \langle x, y \rangle \otimes (\langle x', y' \rangle \otimes \langle x'', y'' \rangle) \tag{17}$$

$$\mathbf{1} \otimes \langle x, y \rangle = \langle 1, 0 \rangle \otimes \langle x, y \rangle \tag{18}$$
$$= \langle 1 \cdot x, 1 \cdot y \rangle \tag{19}$$
$$= \langle x, y \rangle \tag{20}$$
$$= \langle x \cdot 1, y \cdot 1 \rangle \tag{21}$$
$$= \langle x, y \rangle \otimes \mathbf{1} \tag{22}$$

- Multiplication left and right distributes over addition:

$$\langle x, y \rangle \otimes (\langle x', y' \rangle \oplus \langle x'', y'' \rangle) = \langle x, y \rangle \otimes \langle x' + x'', y' + y'' \rangle \tag{23}$$
$$= \langle x \cdot x' + x \cdot x'', x \cdot y' + x \cdot y'' + y \cdot x' + y \cdot x'' \rangle \tag{24}$$
$$= \langle x \cdot x', x \cdot y' + y \cdot x' \rangle \oplus \langle x \cdot x'', x \cdot y'' + y \cdot x'' \rangle \tag{25}$$
$$= (\langle x, y \rangle \otimes \langle x', y' \rangle) \oplus (\langle x, y \rangle \otimes \langle x'', y'' \rangle) \tag{26}$$

$$(\langle x, y \rangle \oplus \langle x', y' \rangle) \otimes \langle x'', y'' \rangle = \langle x + x', y + y' \rangle \otimes \langle x'', y'' \rangle \tag{27}$$
$$= \langle x \cdot x'' + x' \cdot x'', x \cdot y'' + x' \cdot y'' + y \cdot x'' + y' \cdot x'' \rangle \tag{28}$$
$$= \langle x \cdot x'', x \cdot y'' + y \cdot x'' \rangle \oplus \langle x' \cdot x'', x' \cdot y'' + y' \cdot x'' \rangle \tag{29}$$
$$= (\langle x, y \rangle \otimes \langle x'', y'' \rangle) \oplus (\langle x', y' \rangle \otimes \langle x'', y'' \rangle) \tag{30}$$

- Multiplication by $\mathbf{0}$ annihilates $\mathbb{R} \times \mathbb{R}$:

$$\mathbf{0} \otimes \langle x, y \rangle = \langle 0, 0 \rangle \otimes \langle x, y \rangle \tag{31}$$
$$= \langle 0 \cdot x, 0 \cdot y \rangle \tag{32}$$
$$= \langle 0, 0 \rangle \tag{33}$$
$$= \mathbf{0} \tag{34}$$
$$= \langle 0, 0 \rangle \tag{35}$$
$$= \langle x \cdot 0, y \cdot 0 \rangle \tag{36}$$
$$= \langle x, y \rangle \otimes \langle 0, 0 \rangle \tag{37}$$
$$= \langle x, y \rangle \otimes \mathbf{0} \tag{38}$$
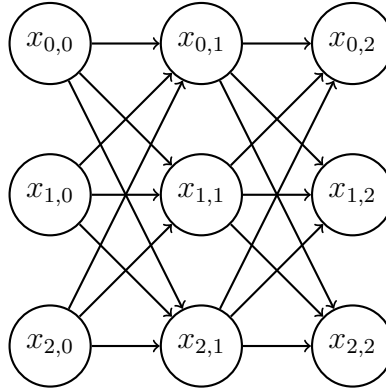
b) Our initial graph looks like Fig. 1.

Figure 1: The initial graph

Where the columns represent the words in $\mathbf{w}$ and the rows represent different tags. We use the algorithm from the script:

Algorithm 1: Forward pass

---
1   $\beta(\mathbf{w}, t_0) = 1$
2   **for** $n = 1 \to N$ **do**
3     $\beta(\mathbf{w}, t_n) = \sum_{t_{n-1} \in \mathcal{T}} \exp(\text{score}_\theta(\langle \langle t_{n-1}, t_n \rangle \rangle, \boldsymbol{w})) \otimes \beta(\mathbf{w}, t_{n-1})$
4   **end**

When we now lift the CRF into the expectation semiring, the forward propagation algorithm changes to:

Algorithm 2: Forward pass

---

1 $\beta(\mathbf{w}, t_0) = \langle 1, 0 \rangle$
2 **for** $n = 1 \to N$ **do**
3 $\quad \beta(\mathbf{w}, t_n) = \oplus_{t_{n-1} \in \mathcal{T}} \langle w, -w \log w \rangle \otimes \beta(\mathbf{w}, t_{n-1})$
4 **end**

Where $w = \exp(\text{score}_\theta(\langle\langle t_n, t_{n+1}\rangle\rangle, \boldsymbol{w}))$.

The output of the forward algorithm lifted into the semiring will yield:

$$\bigoplus_{t_{1:N} \in T^n} \bigotimes_{n=1}^{N} \langle w, -w \log w \rangle \tag{39}$$

We want to show that the result of the forward propagation lifted in the semiring is the same as the unnormalized Entropy:

$$H_u(T_w) = -\sum_{\mathbf{t} \in \mathcal{T}^N} \exp(\text{score}_{\boldsymbol{\theta}}(\mathbf{t}, \boldsymbol{w})) \text{score}_{\boldsymbol{\theta}}(\mathbf{t}, \boldsymbol{w}) \tag{40}$$

$$\tag{41}$$

We show this by induction. Starting with the base case where $N = 1$:

$$\bigoplus_{t_1 \in T^1} \bigotimes_{n=1}^{1} \langle w, -w \log w \rangle = \bigoplus_{t_1 \in T^1} \langle \exp(\text{score}_\theta(\langle t_0, t_1\rangle, \boldsymbol{w})), \tag{42}$$

$$- \exp(\text{score}_\theta(\langle t_0, t_1\rangle, \boldsymbol{w})) \log(\exp(\text{score}_\theta(\langle t_0, t_1\rangle, \boldsymbol{w}))) \rangle \tag{43}$$

$$= \bigoplus_{t \in T} \Big\langle \exp(\text{score}_\theta(t, \boldsymbol{w})), \tag{44}$$

$$- \exp(\text{score}_\theta(t, \boldsymbol{w})) \log(\exp(\text{score}_\theta(t, \boldsymbol{w}))) \Big\rangle \tag{44}$$

$$= \Big\langle \sum_{t \in T} \exp(\text{score}_\theta(t, \boldsymbol{w})), \tag{45}$$

$$- \sum_{t \in T} \exp(\text{score}_\theta(t, \boldsymbol{w})) \log(\exp(\text{score}_\theta(t, \boldsymbol{w}))) \Big\rangle \tag{45}$$

Our induction hypothesis is the following:

$$\bigoplus_{t_{1:i} \in T^i} \bigotimes_{n=1}^{i} \langle w, -w \log w \rangle = \Big\langle \sum_{t \in T^i} \exp(\text{score}_\theta(t, \boldsymbol{w})), - \sum_{t \in T^i} \exp(\text{score}_\theta(t, \boldsymbol{w})) \log(\exp(\text{score}_\theta(t, \boldsymbol{w}))) \Big\rangle \tag{46}$$

Meaning we assume that $\beta(\mathbf{w}, t_i)$ corresponds to the unnormalized entropy of all sequences of lenght i.

3

Now we proceed with the induction step, where $i \to i + 1$:

$$\bigoplus_{t_{1:i+1} \in T^{i+1}} \bigotimes_{n=1}^{i+1} \langle w, -w \log w \rangle$$

$$= \bigoplus_{t_{1:i+1} \in T} \left(\bigoplus_{t_{1:i} \in T^i} \bigotimes_{n=1}^{i} \langle w, -w \log w \rangle \right) \otimes \langle w, -w \log(w) \rangle \tag{47}$$

$$\overset{\text{I.H.}}{=} \bigoplus_{t_{1:i+1} \in T} \left\langle \sum_{t \in T^i} \exp(\text{score}_\theta(t, \boldsymbol{w})), -\sum_{t \in T^i} \exp(\text{score}_\theta((, \boldsymbol{w})t))\text{score}_\theta(t, \boldsymbol{w}) \right\rangle$$

$$\otimes \langle \exp(\text{score}_\theta(\langle t_i, t_{i+1} \rangle, \boldsymbol{w})), -\exp(\text{score}_\theta(\langle t_i, t_{i+1} \rangle, \boldsymbol{w}))\text{score}_\theta(\langle t_i, t_{i+1} \rangle, \boldsymbol{w}) \rangle \tag{48}$$

We will now analyze both parts of the semiring separately:

$$\left(\sum_{t \in T^i} \exp(\text{score}_\theta(t, \boldsymbol{w})) \right) \exp(\text{score}_\theta(\langle t_i, t_{i+1} \rangle, \boldsymbol{w})) \tag{49}$$

$$= \sum_{t \in T^i} \exp(\text{score}_\theta(t, \boldsymbol{w})) + \text{score}_\theta(\langle t_i, t_{i+1} \rangle, \boldsymbol{w}) \tag{50}$$

And the second part:

$$\left(\sum_{t \in T^i} \exp(\text{score}_\theta(t, \boldsymbol{w})) \right) - \exp(\text{score}_\theta(\langle t_i, t_{i+1} \rangle, \boldsymbol{w}))\text{score}_\theta(\langle t_i, t_{i+1} \rangle, \boldsymbol{w})$$

$$+ \left(-\sum_{t \in T^i} \exp(\text{score}_\theta((, \boldsymbol{w})t))\text{score}_\theta(t, \boldsymbol{w}) \right) \exp(\text{score}_\theta(\langle t_i, t_{i+1} \rangle, \boldsymbol{w})) \tag{51}$$

$$= -\sum_{t \in T^i} \exp(\text{score}_\theta(t, \boldsymbol{w}) + \text{score}_\theta(\langle t_i, t_{i+1} \rangle, \boldsymbol{w}))\text{score}_\theta(\langle t_i, t_{i+1} \rangle, \boldsymbol{w})$$

$$- \left(\sum_{t \in T^i} \exp(\text{score}_\theta((, \boldsymbol{w})t))\text{score}_\theta(t, \boldsymbol{w}) \right) \exp(\text{score}_\theta(\langle t_i, t_{i+1} \rangle, \boldsymbol{w})) \tag{52}$$

$$= -\sum_{t \in T^i} \exp(\text{score}_\theta(t, \boldsymbol{w}) + \text{score}_\theta(\langle t_i, t_{i+1} \rangle, \boldsymbol{w}))\text{score}_\theta(\langle t_i, t_{i+1} \rangle, \boldsymbol{w})$$

$$- \sum_{t \in T^i} \exp(\text{score}_\theta(t, \boldsymbol{w}) + \text{score}_\theta(\langle t_i, t_{i+1} \rangle, \boldsymbol{w}))\text{score}_\theta(t, \boldsymbol{w}) \tag{53}$$

$$= -\sum_{t \in T^i} \exp(\text{score}_\theta(t, \boldsymbol{w}) + \text{score}_\theta(\langle t_i, t_{i+1} \rangle, \boldsymbol{w})) \left(\text{score}_\theta(t, \boldsymbol{w}) + \text{score}_\theta(\langle t_i, t_{i+1} \rangle, \boldsymbol{w}) \right)$$

$$\tag{54}$$

We can now combine the two parts to get with eq. (48):

$$\bigoplus_{t_{1:i+1}\in T} \left\langle \sum_{t\in T^i} \exp(\mathrm{score}_\theta(t, \boldsymbol{w})), -\sum_{t\in T^i} \exp(\mathrm{score}_\theta((, \boldsymbol{w})t))\mathrm{score}_\theta(t, \boldsymbol{w}) \right\rangle$$
$$\otimes \left\langle \exp(\mathrm{score}_\theta(\langle t_i, t_{i+1}\rangle, \boldsymbol{w})), -\exp(\mathrm{score}_\theta(\langle t_i, t_{i+1}\rangle, \boldsymbol{w}))\mathrm{score}_\theta(\langle t_i, t_{i+1}\rangle, \boldsymbol{w})\right\rangle \tag{55}$$

$$\overset{\text{def. } \otimes}{=} \bigoplus_{t_{1:i+1}\in T} \left\langle \left(\sum_{t\in T^i} \exp(\mathrm{score}_\theta(t, \boldsymbol{w}))\right)\exp(\mathrm{score}_\theta(\langle t_i, t_{i+1}\rangle, \boldsymbol{w})),\right.$$

$$\left(\sum_{t\in T^i} \exp(\mathrm{score}_\theta(t, \boldsymbol{w}))\right) - \exp(\mathrm{score}_\theta(\langle t_i, t_{i+1}\rangle, \boldsymbol{w}))\mathrm{score}_\theta(\langle t_i, t_{i+1}\rangle, \boldsymbol{w})$$

$$\left. + \left(-\sum_{t\in T^i} \exp(\mathrm{score}_\theta((, \boldsymbol{w})t))\mathrm{score}_\theta(t, \boldsymbol{w})\right)\exp(\mathrm{score}_\theta(\langle t_i, t_{i+1}\rangle, \boldsymbol{w}))\right\rangle \tag{56}$$

$$\overset{eq. (50)}{=} \bigoplus_{t_{1:i+1}\in T} \left\langle \sum_{t\in T^i} \exp(\mathrm{score}_\theta(t, \boldsymbol{w})) + \mathrm{score}_\theta(\langle t_i, t_{i+1}\rangle, \boldsymbol{w}),\right.$$

$$\left(\sum_{t\in T^i} \exp(\mathrm{score}_\theta(t, \boldsymbol{w}))\right) - \exp(\mathrm{score}_\theta(\langle t_i, t_{i+1}\rangle, \boldsymbol{w}))\mathrm{score}_\theta(\langle t_i, t_{i+1}\rangle, \boldsymbol{w})$$

$$\left. + \left(-\sum_{t\in T^i} \exp(\mathrm{score}_\theta((, \boldsymbol{w})t))\mathrm{score}_\theta(t, \boldsymbol{w})\right)\exp(\mathrm{score}_\theta(\langle t_i, t_{i+1}\rangle, \boldsymbol{w}))\right\rangle \tag{57}$$

$$\overset{eq. (54)}{=} \bigoplus_{t_{1:i+1}\in T} \left\langle \sum_{t\in T^i} \exp(\mathrm{score}_\theta(t, \boldsymbol{w})) + \mathrm{score}_\theta(\langle t_i, t_{i+1}\rangle, \boldsymbol{w}),\right.$$

$$\left. -\sum_{t\in T^i} \exp(\mathrm{score}_\theta(t, \boldsymbol{w}) + \mathrm{score}_\theta(\langle t_i, t_{i+1}\rangle, \boldsymbol{w}))\left(\mathrm{score}_\theta(t, \boldsymbol{w}) + \mathrm{score}_\theta(\langle t_i, t_{i+1}\rangle, \boldsymbol{w})\right)\right\rangle$$
$$\tag{58}$$

$$\overset{\text{def. } \oplus}{=} \left\langle \sum_{t_{i+1}\in T}\sum_{t\in T^i} \exp(\mathrm{score}_\theta(t, \boldsymbol{w})) + \mathrm{score}_\theta(\langle t_i, t_{i+1}\rangle, \boldsymbol{w}),\right.$$

$$\left. \sum_{t_{i+1}\in T}\sum_{t\in T^i} \exp(\mathrm{score}_\theta(t, \boldsymbol{w}) + \mathrm{score}_\theta(\langle t_i, t_{i+1}\rangle, \boldsymbol{w}))\left(\mathrm{score}_\theta(t, \boldsymbol{w}) + \mathrm{score}_\theta(\langle t_i, t_{i+1}\rangle, \boldsymbol{w})\right)\right\rangle$$
$$\tag{59}$$

$$= \left\langle \sum_{t\in T^{i+1}} \exp(\mathrm{score}_\theta(t, \boldsymbol{w})), -\sum_{t\in T^{i+1}} \exp(\mathrm{score}_\theta(t, \boldsymbol{w}))\mathrm{score}_\theta(t, \boldsymbol{w}) \right\rangle \tag{60}$$

Which concludes our induction proof and we have shown that eq. (40) holds.

c) We want to prove:

$$H(T_w) = Z(\boldsymbol{w})^{-1}H_U(T) + \log(Z(\boldsymbol{w}) \tag{61}$$

5

$$H(T_w) = -\sum_{t \in T^N} p(t \mid w) \cdot \log(p(t \mid w)) \qquad \text{(def. H)}$$

$$\text{(62)}$$

$$= -\sum_{t \in T^N} \frac{\exp(\text{score}_\theta(t, \boldsymbol{w}))}{Z(\boldsymbol{w})} \log \left(\frac{\exp(\text{score}_\theta(t, \boldsymbol{w}))}{\sum_{t' \in T^N} \exp(\text{score}_\theta(t', \boldsymbol{w}))} \right) \qquad \text{(def. p)}$$

$$\text{(63)}$$

$$= -\sum_{t \in T^N} \frac{\exp(\text{score}_\theta(t, \boldsymbol{w}))}{Z(\boldsymbol{w})} \log \left(\frac{\exp(\text{score}_\theta(t, \boldsymbol{w}))}{Z(\boldsymbol{w})} \right) \qquad \text{(64)}$$

$$= -\sum_{t \in T^N} \frac{\exp(\text{score}_\theta(t, \boldsymbol{w}))}{Z(\boldsymbol{w})} (\text{score}_\theta(t, \boldsymbol{w}) - \log Z(\boldsymbol{w})) \qquad \text{(65)}$$

$$= -\sum_{t \in T^N} \frac{\exp(\text{score}_\theta(t, \boldsymbol{w}))\text{score}_\theta(t, \boldsymbol{w}) - \exp(\text{score}_\theta(t, \boldsymbol{w})) \log Z(\boldsymbol{w})}{Z(\boldsymbol{w})} \qquad \text{(66)}$$

$$= -\sum_{t \in T^N} \frac{\exp(\text{score}_\theta(t, \boldsymbol{w}))\text{score}_\theta(t, \boldsymbol{w})}{Z(\boldsymbol{w})} + \sum_{t \in T^N} \frac{\exp(\text{score}_\theta(t, \boldsymbol{w})) \log Z(\boldsymbol{w})}{Z(\boldsymbol{w})}$$

$$\text{(67)}$$

$$= H_U(T_{\boldsymbol{w}})Z(\boldsymbol{w})^{-1} + \frac{\log(Z(\boldsymbol{w}))}{Z(\boldsymbol{w})} \sum_{t \in T^N} \exp(\text{score}_\theta(t, \boldsymbol{w})) \qquad \text{(68)}$$

$$= H_U(T_{\boldsymbol{w}})Z(\boldsymbol{w})^{-1} + \log(Z(\boldsymbol{w})) \qquad \text{(69)}$$

$$\text{(70)}$$