Simon Wachter: Assignment 2

siwachte@ethz.ch, 19-920-198

27/11/2022 - 16:46h

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 $\mathbf{0}$:

$$\begin{aligned}
(\langle x, y\rangle \oplus \langle x', y'\rangle) \oplus \langle x'', y''\rangle &= \langle x + x', y + y'\rangle \oplus \langle x'', y''\rangle && (1)\\
&= \langle x + x' + x'', y + y' + y''\rangle && (2)\\
&= \langle x, y\rangle \oplus \langle x' + x'', y' + y''\rangle && (3)\\
&= \langle x, y\rangle \oplus (\langle x', y'\rangle \oplus \langle x'', y''\rangle) && (4)
\end{aligned}$$

$$\begin{aligned}
\mathbf{0} + \langle x, y\rangle &= \langle 0, 0\rangle \oplus \langle x, y\rangle && (5)\\
&= \langle 0 + x, 0 + y\rangle && (6)\\
&= \langle x, y\rangle && (7)\\
&= \langle x + 0, y + 0\rangle && (8)\\
&= \langle x, y\rangle + \mathbf{0} && (9)
\end{aligned}$$

$$\begin{aligned}
\langle x, y\rangle + \langle x', y'\rangle &= \langle x + x', y + y'\rangle && (10)\\
&= \langle x' + x, y' + y\rangle && (11)\\
&= \langle x', y'\rangle + \langle x, y\rangle && (12)
\end{aligned}$$

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

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

$$\begin{aligned}
\mathbf{1} \otimes \langle x, y\rangle &= \langle 1, 0\rangle \otimes \langle x, y\rangle && (18)\\
&= \langle 1 \cdot x, 1 \cdot y\rangle && (19)\\
&= \langle x, y\rangle && (20)\\
&= \langle x \cdot 1, y \cdot 1\rangle && (21)\\
&= \langle x, y\rangle \otimes \mathbf{1} && (22)
\end{aligned}$$

- Multiplication left and right distributes over addition:

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

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

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

$$\begin{aligned}
\mathbf{0} \otimes \langle x, y\rangle &= \langle 0, 0\rangle \otimes \langle x, y\rangle && (31)\\
&= \langle 0 \cdot x, 0 \cdot y\rangle && (32)\\
&= \langle 0, 0\rangle && (33)\\
&= \mathbf{0} && (34)\\
&= \langle 0, 0\rangle && (35)\\
&= \langle x \cdot 0, y \cdot 0\rangle && (36)\\
&= \langle x, y\rangle \otimes \langle 0, 0\rangle && (37)\\
&= \langle x, y\rangle \otimes \mathbf{0} && (38)
\end{aligned}$$

b) Our initial graph looks like Fig. 1.

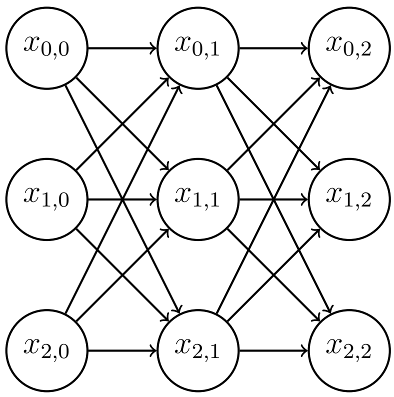

Figure 1: The initial graph

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

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

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

Algorithm 2: Forward pass

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

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

$$\tag{41}$$

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

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

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

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

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

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

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

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.

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

$$\bigoplus_{t_{1:N}\in T^n} \bigotimes_{n=1}^{N} \langle w, -w \log w\rangle \tag{47}$$

$$= \bigoplus_{t_{1:N-1}\in T^{N-1}} \bigoplus_{t_n \in T} \bigotimes_{n=1}^{N} \langle w, -w \log w\rangle \tag{48}$$

$$\begin{aligned}
&= \bigoplus_{t_{1:N-1}\in T^{N-1}} \bigotimes_{n=1}^{N-1} \langle \exp(\mathrm{score}_\theta(\langle t_{n-1}, t_n\rangle, \boldsymbol{w})), -\exp(\mathrm{score}_\theta(\langle t_{n-1}, t_n\rangle, \boldsymbol{w})) \log(\exp(\mathrm{score}_\theta(\langle t_{n-1}, t_n\rangle, \boldsymbol{w} \\
&\quad \otimes \bigoplus_{t_N \in T} \langle \exp(\mathrm{score}_\theta(\langle t_{N-1}, t_N\rangle, \boldsymbol{w})), -\exp(\mathrm{score}_\theta(\langle t_{N-1}, t_N\rangle, \boldsymbol{w})) \log(\exp(\mathrm{score}_\theta(\langle t_{N-1}, t_N\rangle, \boldsymbol{w})))
\end{aligned} \tag{49}$$

$$\begin{aligned}
&= \bigoplus_{t_1 \in T} \langle \exp(\mathrm{score}_\theta(\langle t_0, t_1\rangle, \boldsymbol{w})), -\exp(\mathrm{score}_\theta(\langle t_0, t_1\rangle, \boldsymbol{w})) \log(\exp(\mathrm{score}_\theta(\langle t_0, t_1\rangle, \boldsymbol{w})))\rangle \\
&\quad \otimes \bigoplus_{t_2 \in T} \langle \exp(\mathrm{score}_\theta(\langle t_1, t_2\rangle, \boldsymbol{w})), -\exp(\mathrm{score}_\theta(\langle t_1, t_2\rangle, \boldsymbol{w})) \log(\exp(\mathrm{score}_\theta(\langle t_1, t_2\rangle, \boldsymbol{w})))\rangle \\
&\quad \otimes \cdots \\
&\quad \otimes \bigoplus_{t_N \in T} \langle \exp(\mathrm{score}_\theta(\langle t_{N-1}, t_N\rangle, \boldsymbol{w})), -\exp(\mathrm{score}_\theta(\langle t_{N-1}, t_N\rangle, \boldsymbol{w})) \log(\exp(\mathrm{score}_\theta(\langle t_{N-1}, t_N\rangle, \boldsymbol{w})))
\end{aligned}$$

$$= \left\langle \sum_{\mathbf{t}\in\mathcal{T}^N} \exp(score_{\boldsymbol{\theta}}(\mathbf{t}, \boldsymbol{w})), -\sum_{\mathbf{t}\in\mathcal{T}^N} \exp(score_{\boldsymbol{\theta}}(\mathbf{t}, \boldsymbol{w})) score_{\boldsymbol{\theta}}(\mathbf{t}, \boldsymbol{w}) \right\rangle \tag{50}$$

c) We want to prove:

$$H(T_w) = Z(\boldsymbol{w})^{-1} H_U(T) + \log(Z(\boldsymbol{w}) \tag{51}$$

$$
\begin{aligned}
H(T_w) &= -\sum_{t \in T^N} p(t \mid w) \cdot \log(p(t \mid w)) && \text{(def. H)} \\
& && (52) \\
&= -\sum_{t \in T^N} \frac{\exp(\mathrm{score}_\theta(t, \boldsymbol{w}))}{Z(\boldsymbol{w})} \log\left(\frac{\exp(\mathrm{score}_\theta(t, \boldsymbol{w}))}{\sum_{t' \in T^N} \exp(\mathrm{score}_\theta(t', \boldsymbol{w}))}\right) && \text{(def. p)} \\
& && (53) \\
&= -\sum_{t \in T^N} \frac{\exp(\mathrm{score}_\theta(t, \boldsymbol{w}))}{Z(\boldsymbol{w})} \log\left(\frac{\exp(\mathrm{score}_\theta(t, \boldsymbol{w}))}{Z(\boldsymbol{w})}\right) && (54) \\
&= -\sum_{t \in T^N} \frac{\exp(\mathrm{score}_\theta(t, \boldsymbol{w}))}{Z(\boldsymbol{w})} (\mathrm{score}_\theta(t, \boldsymbol{w}) - \log Z(\boldsymbol{w})) && (55) \\
&= -\sum_{t \in T^N} \frac{\exp(\mathrm{score}_\theta(t, \boldsymbol{w}))\mathrm{score}_\theta(t, \boldsymbol{w}) - \exp(\mathrm{score}_\theta(t, \boldsymbol{w})) \log Z(\boldsymbol{w}))}{Z(\boldsymbol{w})} && (56) \\
&= -\sum_{t \in T^N} \frac{\exp(\mathrm{score}_\theta(t, \boldsymbol{w}))\mathrm{score}_\theta(t, \boldsymbol{w})}{Z(\boldsymbol{w})} + \sum_{t \in T^N} \frac{\exp(\mathrm{score}_\theta(t, \boldsymbol{w})) \log Z(\boldsymbol{w}))}{Z(\boldsymbol{w})} \\
& && (57) \\
&= H_U(T_{\boldsymbol{w}}) Z(\boldsymbol{w})^{-1} + \frac{\log(Z(\boldsymbol{w}))}{Z(\boldsymbol{w})} \sum_{t \in T^N} \exp(\mathrm{score}_\theta(t, \boldsymbol{w})) && (58) \\
&= H_U(T_{\boldsymbol{w}}) Z(\boldsymbol{w})^{-1} + \log(Z(\boldsymbol{w})) && (59) \\
& && (60)
\end{aligned}
$$