Prof. Ryan Cotterell

Yannick Wattenberg: Assignment 02

ywattenberg@inf.ethz.ch, 19-947-464.

25/11/2022 - 15:09h

1 Question: Entropy of a Conditional Random Field

a)

Prove that the **expectation semiring** satisfies the semiring axioms:
Let $x_1, y_1, x_2, y_2, x_3, y_3 \in \mathbb{R}$ for this subsection.

Axiom 1

$$(\mathbb{R} \times \mathbb{R}, \oplus, \langle 0, 0\rangle) \text{ is a commutative monoid with identity element } \langle 0, 0\rangle$$

Associativity and Commutativity of $\oplus$:

$$\begin{aligned}
(\langle x_1, y_1\rangle \oplus \langle x_2, y_2\rangle) \oplus \langle x_3, y_3\rangle &\stackrel{!}{=} \langle x_1, y_1\rangle \oplus (\langle x_3, y_3\rangle \oplus \langle x_2, y_2\rangle) && (1)\\
(\langle x_1, y_1\rangle \oplus \langle x_2, y_2\rangle) \oplus \langle x_3, y_3\rangle &= \langle x_1 + x_2, y_1 + y_2\rangle \oplus \langle x_3, y_3\rangle && \text{(def. } \oplus) \quad (2)\\
&= \langle (x_1 + x_2) + x_3, (y_1 + y_2) + y_3\rangle && \text{(def. } \oplus) \quad (3)\\
&= \langle x_1 + (x_2 + x_3), y_1 + (y_2 + y_3)\rangle && \text{(ass. of } +) \quad (4)\\
&= \langle x_1, y_1\rangle \oplus \langle x_2 + x_3, y_2 + y_3\rangle && \text{(def. } \oplus) \quad (5)\\
&= \langle x_1, y_1\rangle \oplus \langle x_3 + x_2, y_3 + y_2\rangle && \text{(comm. of } +) \quad (6)\\
&= \langle x_1, y_1\rangle \oplus (\langle x_3, y_3\rangle \oplus \langle x_2, y_2\rangle) && \text{(def. } \oplus) \quad (7)
\end{aligned}$$

$\langle 0, 0\rangle$ is the identity element:

$$\begin{aligned}
\langle 0, 0\rangle \oplus \langle x_1, y_1\rangle &\stackrel{!}{=} \langle x_1, y_1\rangle && (8)\\
\langle 0, 0\rangle \oplus \langle x_1, y_1\rangle &= \langle x_1 + 0, y_1 + 0\rangle && \text{(def. } \oplus) \quad (9)\\
&= \langle x_1, y_1\rangle && (10)\\
&= \langle x_1 + 0, y_1 + 0\rangle && (11)\\
&= \langle x_1, y_1\rangle \oplus \langle 0, 0\rangle && \text{(def. } \oplus) \quad (12)
\end{aligned}$$

Axiom 2

$$(\mathbb{R}\times\mathbb{R},\otimes,\langle 1,0\rangle) \text{ is a monoid with identity element } \langle 1,0\rangle$$

Associativity of $\otimes$:

$$\begin{aligned}
(\langle x_1,y_1\rangle\otimes\langle x_2,y_2\rangle)\otimes\langle x_3,y_3\rangle &\overset{!}{=} \langle x_1,y_1\rangle\otimes(\langle x_2,y_2\rangle\otimes\langle x_3,y_3\rangle) && && (13)\\
(\langle x_1,y_1\rangle\otimes\langle x_2,y_2\rangle)\otimes\langle x_3,y_3\rangle &= \langle x_1\cdot x_2, x_1\cdot y_2+y_1\cdot x_2\rangle\otimes\langle x_3,y_3\rangle && \text{(def. }\otimes) && (14)\\
&=\langle (x_1\cdot x_2)\cdot x_3, (x_1\cdot x_2)\cdot y_3+(x_1\cdot y_2+y_1\cdot x_2)\cdot x_3\rangle && \text{(def. }\otimes) && (15)\\
&=\langle (x_1\cdot x_2)\cdot x_3, x_1\cdot x_2\cdot y_3+x_1\cdot y_2\cdot x_3+y_1\cdot x_2\cdot x_3\rangle && \text{(diss. }\cdot) && (16)\\
&=\langle (x_1\cdot x_2)\cdot x_3, x_1\cdot (x_2\cdot y_3+y_2\cdot x_3)+y_1\cdot x_2\cdot x_3\rangle && \text{(diss. }\cdot) && (17)\\
&=\langle x_1\cdot (x_2\cdot x_3), x_1\cdot (x_2\cdot y_3+y_2\cdot x_3)+y_1\cdot (x_2\cdot x_3)\rangle && \text{(ass. }\cdot) && (18)\\
&=\langle x_1,y_1\rangle\otimes\langle x_2\cdot x_3, x_2\cdot y_3+y_2\cdot x_3\rangle && \text{(def. }\otimes) && (19)\\
&=\langle x_1,y_1\rangle\otimes(\langle x_2,y_2\rangle\otimes\langle x_3,y_3\rangle) && \text{(def. }\otimes) && (20)
\end{aligned}$$

$\langle 1,0\rangle$ is the identity element:

$$\begin{aligned}
\langle x_1,y_1\rangle\otimes\langle 1,0\rangle &\overset{!}{=} \langle x_1,y_1\rangle && && (21)\\
\langle x_1,y_1\rangle\otimes\langle 1,0\rangle &= \langle x_1\cdot 1, x_1\cdot 0+y_1\cdot 1\rangle && \text{(def. }\otimes) && (22)\\
&= \langle x_1,y_1\rangle && && (23)\\
&= \langle 1\cdot x_1, 1\cdot y_1+0\cdot x_1\rangle && && (24)\\
&= \langle 1,0\rangle\otimes\langle x_1,y_1\rangle && \text{(def. }\otimes) && (25)
\end{aligned}$$

Axiom 3

$\otimes$ distributes left and right over $\oplus$:

$$\langle x_1,y_1\rangle\otimes(\langle x_2,y_2\rangle\oplus\langle x_3,y_3\rangle) \overset{!}{=} (\langle x_1,y_1\rangle\otimes\langle x_2,y_2\rangle)\oplus(\langle x_1,y_1\rangle\otimes\langle x_3,y_3\rangle) \qquad (26)$$

$$\begin{aligned}
\langle x_1,y_1\rangle\otimes(\langle x_2,y_2\rangle\oplus\langle x_3,y_3\rangle) &= \langle x_1,y_1\rangle\otimes\langle x_2+x_3, y_2+y_3\rangle && \text{(def. }\oplus) && (27)\\
&=\langle x_1\cdot(x_2+x_3), x_1\cdot(x_2+x_3)+y_1\cdot(y_2+y_3)\rangle && \text{(def. }\otimes) && (28)\\
&=\langle (x_1\cdot x_2)+(x_1\cdot x_3), (x_1\cdot y_2)+(y_1\cdot x_2)+(x_1\cdot y_3)+(y_1\cdot x_3)\rangle && \text{(diss. }\cdot) && (29)\\
&=\langle x_1\cdot x_2, (x_1\cdot y_2)+(y_1\cdot x_2)\rangle\oplus\langle x_1\cdot x_3, (x_1\cdot y_3)+(y_1\cdot x_3)\rangle && \text{(def. }\oplus) && (30)\\
&=(\langle x_1,y_1\rangle\otimes\langle x_2,y_2\rangle)\oplus(\langle x_1,y_1\rangle\otimes\langle x_3,y_3\rangle) && \text{(def. }\otimes) && (31)
\end{aligned}$$

$$(\langle x_2,y_2\rangle\oplus\langle x_3,y_3\rangle)\otimes\langle x_1,y_1\rangle \overset{!}{=} (\langle x_2,y_2\rangle\otimes\langle x_1,y_1\rangle)\oplus(\langle x_3,y_3\rangle\otimes\langle x_1,y_1\rangle) \qquad (32)$$

$$\begin{aligned}
(\langle x_2,y_2\rangle\oplus\langle x_3,y_3\rangle)\otimes\langle x_1,y_1\rangle &= (\langle x_2+x_3, y_2+y_3\rangle)\otimes\langle x_1,y_1\rangle && \text{(def. }\oplus) && (33)\\
&=\langle (x_2+x_3)\cdot x_1, (x_2+x_3)\cdot y_1+(y_2+y_3)\cdot x_1\rangle && \text{(def. }\otimes) && (34)\\
&=\langle (x_2\cdot x_1)\cdot(x_3\cdot x_1), (x_2\cdot y_1)+(x_3\cdot y_1)+(y_2\cdot x_1)+(y_3\cdot x_1)\rangle && \text{(diss. }\cdot) && (35)\\
&=\langle x_2\cdot x_1, x_2\cdot y_1+y_2\cdot x_1\rangle\oplus\langle x_3\cdot x_1, x_3\cdot y_1+y_3\cdot x_1\rangle && \text{(def. }\oplus) && (36)\\
&=(\langle x_2,y_2\rangle\otimes\langle x_1,y_1\rangle)\oplus(\langle x_3,y_3\rangle\otimes\langle x_1,y_1\rangle) && \text{(def. }\otimes) && (37)
\end{aligned}$$

Axiom 4

$$\langle 0,0\rangle \otimes \langle x_1,y_1\rangle \overset{!}{=} \langle 0,0\rangle \overset{!}{=} \langle x_1,y_1\rangle \otimes \langle 0,0\rangle \tag{38}$$
$$\langle 0,0\rangle \otimes \langle x_1,y_1\rangle = \langle 0\cdot x_1, 0\cdot y_1 + 0\cdot x_1\rangle \qquad \text{(def. } \otimes) \tag{39}$$
$$= \langle 0,0\rangle \tag{40}$$
$$= \langle x_1\cdot 0, x_1\cdot 0 + y_1\cdot 0\rangle \tag{41}$$
$$= \langle x_1,y_1\rangle \otimes \langle 0,0\rangle \qquad \text{(def. } \otimes) \tag{42}$$

With this, we have proven that the **expectation semiring** is a semiring.

b)

The definition of the forward algorithm is as follows: Where $\forall t_i \in T$ is implicit. We raise this

---

Algorithm 1: Forward algorithm

$\beta(\mathbf{w}, t_0) = \mathbb{1}$
for $i = 1$ **to** N **do**
  $\beta(\mathbf{w}, t_i) = \oplus_{t_{i-1}\in T} exp(score(t_{i-1}, t_i, \mathbf{w})) \otimes \beta(\mathbf{w}, t_{i-1})$
end

---

into the expectation semiring by replacing the $\oplus$ and $\otimes$ with the corresponding operations in the expectation semiring. We also replace the initialization of $\beta(w, t_0)$ with the neutral element of multiplication in the expectation semiring. Which yields the following algorithm:

Algorithm 2: Forward algorithm in the expectation semiring

$\beta(\mathbf{w}, t_0) = \langle 1, 0\rangle$
for $i = 1$ **to** N **do**
  $w = exp(score(t_{i-1}, t_i, \mathbf{w}))$
  $\beta(\mathbf{w}, t_i) = \oplus_{t_{i-1}\in T} \langle w, -w \cdot log w\rangle \otimes \beta(\mathbf{w}, t_{i-1})$
end

We want to prove that the result of the forward algorithm lifted in the expectation semiring computes the unnormalized Entropy defined as:

$$H_u(T_{\mathbf{w}}) = -\sum_{t\in T^N} exp(score_\theta(t, \mathbf{w})) \cdot score_\theta(t, \mathbf{w}) \tag{43}$$

$$\tag{44}$$

The lifted forward algorithm is then equal to:

$$\bigoplus_{t_{1:N}\in T^N} \bigotimes_{n=1}^{N} \langle w, -w \cdot \log(w)\rangle \tag{45}$$

We prove the statement by induction on the length of the sequence N. The base case is trivial, since the forward algorithm only computes the score of the first token:

$$\bigoplus_{t_{1:1}\in T^1}\bigotimes_{n=1}^{1}\langle w, -w\cdot\log(w)\rangle \tag{46}$$

$$=\bigoplus_{t_1\in T}\langle w, -w\cdot\log(w)\rangle \tag{47}$$

$$(\text{def. w}) = \bigoplus_{t_1\in T}\langle \exp(\mathit{score}(t_0, t_1, \mathbf{w})), -\exp(\mathit{score}(t_0, t_1, \mathbf{w}))\cdot \mathit{score}(t_0, t_1, \mathbf{w})\rangle \tag{48}$$

$$=\bigoplus_{t\in T}\langle \exp(\mathit{score}_\theta(t, \mathbf{w})), -\mathit{exp}(\mathit{score}_\theta(t, \mathbf{w}))\cdot \mathit{score}_\theta(t, \mathbf{w})\rangle \tag{49}$$

$$(\text{def. }\oplus) = \left\langle \sum_{t\in T^1}\exp(\mathit{score}_\theta(t, \mathbf{w})), -\sum_{t\in T^1}\mathit{exp}(\mathit{score}_\theta(t, \mathbf{w}))\cdot \mathit{score}_\theta(t, \mathbf{w})\right\rangle \tag{50}$$

Our induction hypothesis is that the forward algorithm computes the unnormalized entropy for sequences of length i:

$$\bigoplus_{t_{1:i}\in T^i}\bigotimes_{n=1}^{i}\langle w, -w\cdot\log(w)\rangle = \left\langle \sum_{t\in T^i}\mathit{exp}(\mathit{score}_\theta(t, \mathbf{w})), -\sum_{t\in T^i}\mathit{exp}(\mathit{score}_\theta(t, \mathbf{w}))\cdot \mathit{score}_\theta(t, \mathbf{w})\right\rangle$$

Induction step ($i \to i+1$):

$$\bigoplus_{t_{1:i+1}\in T^{i+1}}\bigotimes_{n=1}^{i+1}\langle w, -w\cdot\log(w)\rangle \tag{51}$$

$$=\bigoplus_{t_{i+1}\in T}\left(\bigoplus_{t_{1:i}\in T^i}\bigotimes_{n=1}^{i}\langle w, -w\cdot\log(w)\rangle\right)\otimes\langle w, -w\log(w)\rangle \tag{52}$$

$$(\text{def. IH}) = \bigoplus_{t_{i+1}\in T}\left\langle \sum_{t\in T^i}\mathit{exp}(\mathit{score}_\theta(t, \mathbf{w})), -\sum_{t\in T^i}\mathit{exp}(\mathit{score}_\theta(t, \mathbf{w}))\cdot \mathit{score}_\theta(t, \mathbf{w})\right\rangle \tag{53}$$

$$\otimes\langle w, -w\log(w)\rangle \tag{54}$$

$$(\text{def. w}) = \bigoplus_{t_{i+1}\in T}\left\langle \sum_{t\in T^i}\mathit{exp}(\mathit{score}_\theta(t, \mathbf{w})), -\sum_{t\in T^i}\mathit{exp}(\mathit{score}_\theta(t, \mathbf{w}))\cdot \mathit{score}_\theta(t, \mathbf{w})\right\rangle \tag{55}$$

$$\otimes\langle \exp(\mathit{score}(t_i, t_{i+1}, \mathbf{w})), -\exp(\mathit{score}(t_i, t_{i+1}, \mathbf{w}))\cdot \mathit{score}(t_i, t_{i+1}, \mathbf{w})\rangle \tag{56}$$

$$\left(\sum_{t\in T^i}\mathit{exp}(\mathit{score}_\theta(t, \mathbf{w}))\right)\cdot\exp(\mathit{score}(t_i, t_{i+1}, \mathbf{w})) \tag{57}$$

$$=\sum_{t\in T^i}\mathit{exp}(\mathit{score}_\theta(t, \mathbf{w}) + \mathit{score}(t_i, t_{i+1}, \mathbf{w})) \tag{58}$$

Further we also have:

$$\left(\sum_{t\in T^i} exp(score_\theta(t,\mathbf{w}))\right)\cdot(-\exp(score(t_i,t_{i+1},\mathbf{w}))\cdot score(t_i,t_{i+1},\mathbf{w})) \tag{59}$$

$$+\left(-\sum_{t\in T^i} exp(score_\theta(t,\mathbf{w}))\cdot score_\theta(t,\mathbf{w})\right)\cdot\exp(score(t_i,t_{i+1},\mathbf{w})) \tag{60}$$

$$=-\sum_{t\in T^i} exp(score_\theta(t,\mathbf{w})+score(t_i,t_{i+1},\mathbf{w}))\cdot score(t_i,t_{i+1},\mathbf{w}) \tag{61}$$

$$-\sum_{t\in T^i} exp(score_\theta(t,\mathbf{w})+score(t_i,t_{i+1},\mathbf{w}))\cdot score_\theta(t,\mathbf{w}) \tag{62}$$

$$=-\sum_{t\in T^i} exp(score_\theta(t,\mathbf{w})+score(t_i,t_{i+1},\mathbf{w}))\cdot score_\theta(t,\mathbf{w})\cdot score(t_i,t_{i+1},\mathbf{w}) \tag{63}$$

Using the equalities from equation 57 to equation 57 we can rewrite the induction step as:

$$56\Rightarrow\bigoplus_{t_{i+1}\in T}\left\langle\left(\sum_{t\in T^i} exp(score_\theta(t,\mathbf{w}))\right)\cdot\exp(score(t_i,t_{i+1},\mathbf{w})),\right. \tag{64}$$

$$\left(\sum_{t\in T^i} exp(score_\theta(t,\mathbf{w}))\right)\cdot(-\exp(score(t_i,t_{i+1},\mathbf{w}))\cdot score(t_i,t_{i+1},\mathbf{w})) \tag{65}$$

$$\left.+\left(-\sum_{t\in T^i} exp(score_\theta(t,\mathbf{w}))\cdot score_\theta(t,\mathbf{w})\right)\cdot\exp(score(t_i,t_{i+1},\mathbf{w}))\right\rangle \tag{66}$$

$$(\text{with } 63)=\bigoplus_{t_{i+1}\in T}\left\langle\left(\sum_{t\in T^i} exp(score_\theta(t,\mathbf{w}))\right)\cdot\exp(score(t_i,t_{i+1},\mathbf{w})),\right. \tag{67}$$

$$\left.-\sum_{t\in T^i} exp(score_\theta(t,\mathbf{w})+score(t_i,t_{i+1},\mathbf{w}))\cdot score_\theta(t,\mathbf{w})\cdot score(t_i,t_{i+1},\mathbf{w})\right\rangle \tag{68}$$

$$(\text{with } 57)=\bigoplus_{t_{i+1}\in T}\left\langle\sum_{t\in T^i} exp(score_\theta(t,\mathbf{w})+score(t_i,t_{i+1},\mathbf{w})),\right. \tag{69}$$

$$\left.-\sum_{t\in T^i} exp(score_\theta(t,\mathbf{w})+score(t_i,t_{i+1},\mathbf{w}))\cdot score_\theta(t,\mathbf{w})\cdot score(t_i,t_{i+1},\mathbf{w})\right\rangle \tag{70}$$

$$(\text{def. }\oplus)=\left\langle\sum_{t_{i+1}\in T}\sum_{t\in T^i} exp(score_\theta(t,\mathbf{w})+score(t_i,t_{i+1},\mathbf{w})),\right. \tag{71}$$

$$\left.-\sum_{t_{i+1}\in T}\sum_{t\in T^i} exp(score_\theta(t,\mathbf{w})+score(t_i,t_{i+1},\mathbf{w}))\cdot score_\theta(t,\mathbf{w})\cdot score(t_i,t_{i+1},\mathbf{w})\right\rangle \tag{72}$$

$$=\left\langle\sum_{t\in T^{i+1}} exp(score_\theta(t,\mathbf{w})),-\sum_{t\in T^{i+1}} exp(score_\theta(t,\mathbf{w}))\cdot score_\theta(t,\mathbf{w})\right\rangle \tag{73}$$

Concluding the induction.
From this, we can follow that:

$$\bigoplus_{t_{1:N} \in T^N} \bigotimes_{n=1}^{N} \langle w, -w \cdot \log(w) \rangle = \left\langle \sum_{t \in T^N} exp(score_\theta(t, \mathbf{w})), - \sum_{t \in T^N} exp(score_\theta(t, \mathbf{w})) \cdot score_\theta(t, \mathbf{w}) \right\rangle$$

This shows that the second part of the pair is equal to the unnormalized entropy.

c)

To prove:

$$H(T_w) \overset{!}{=} Z(w)^{-1} H_U(T_w) + \log Z(w) \tag{74}$$

$$H(T_w) = - \sum_{t \in T^N} p(t|w) \cdot \log p(t|w) \qquad \text{(def. } H) \tag{75}$$

$$= - \sum_{t \in T^N} \frac{exp(score_\theta(t, w))}{Z(w)} \cdot \log(\frac{exp(score_\theta(t, w))}{Z(w)}) \qquad \text{(def. } p) \tag{76}$$

$$= - \sum_{t \in T^N} \frac{exp(score_\theta(t, w))}{Z(w)} \cdot (score_\theta(t, w) - \log(Z(w))) \tag{77}$$

$$= - \sum_{t \in T^N} \frac{exp(score_\theta(t, w)) \cdot (score_\theta(t, w) - \log(Z(w)))}{Z(w)} \tag{78}$$

$$= \sum_{t \in T^N} \frac{exp(score_\theta(t, w)) \cdot (-score_\theta(t, w) + \log(Z(w)))}{Z(w)} \tag{79}$$

$$= \sum_{t \in T^N} \frac{exp(score_\theta(t, w)) \cdot \log(Z(w))}{Z(w)} \tag{80}$$

$$- \sum_{t \in T^N} \frac{exp(score_\theta(t, w)) \cdot score_\theta(t, w)}{Z(w)} \tag{81}$$

$$= \sum_{t \in T^N} \frac{exp(score_\theta(t, w)) \cdot \log(Z(w))}{Z(w)} \tag{82}$$

$$- \sum_{t \in T^N} (exp(score_\theta(t, w)) \cdot score_\theta(t, w)) \cdot Z(w)^{-1} \tag{83}$$

$$= H_U(T_w) \cdot Z(w)^{-1} + \sum_{t \in T^N} \frac{exp(score_\theta(t, w)) \cdot \log(Z(w))}{Z(w)} \qquad \text{(def. } H_U) \tag{84}$$

$$= H_U(T_w) \cdot Z(w)^{-1} + \frac{\log(Z(w))}{Z(w)} \cdot \sum_{t \in T^N} exp(score_\theta(t, w)) \tag{85}$$

$$= H_U(T_w) \cdot Z(w)^{-1} + \frac{\log(Z(w))}{Z(w)} \cdot Z(w) \qquad \text{(def. } Z(w)) \tag{86}$$

$$= Z(w)^{-1} H_U(T_w) + \log Z(w) \tag{87}$$