# Yannick Wattenberg: Assignment 02

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

09/12/2022 - 09:15h

## 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

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

Associativity and Commutativity of $\oplus$:

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

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

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

**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 $\otimes$:

$$(\langle x_1, y_1\rangle \otimes \langle x_2, y_2\rangle) \otimes \langle x_3, y_3\rangle \overset{!}{=} \langle x_1, y_1\rangle \otimes (\langle x_2, y_2\rangle \otimes \langle x_3, y_3\rangle) \tag{13}$$

$$(\langle x_1, y_1\rangle \otimes \langle x_2, y_2\rangle) \otimes \langle x_3, y_3\rangle = \langle x_1 \cdot x_2, x_1 \cdot y_2 + y_1 \cdot x_2\rangle \otimes \langle x_3, y_3\rangle \quad \text{(def. } \otimes) \tag{14}$$

$$= \langle (x_1 \cdot x_2) \cdot x_3, (x_1 \cdot x_2) \cdot y_3 + (x_1 \cdot y_2 + y_1 \cdot x_2) \cdot x_3\rangle \quad \text{(def. } \otimes) \tag{15}$$

$$= \langle (x_1 \cdot x_2) \cdot x_3, x_1 \cdot x_2 \cdot y_3 + x_1 \cdot y_2 \cdot x_3 + y_1 \cdot x_2 \cdot x_3\rangle \quad \text{(diss. } \cdot) \tag{16}$$

$$= \langle (x_1 \cdot x_2) \cdot x_3, x_1 \cdot (x_2 \cdot y_3 + y_2 \cdot x_3) + y_1 \cdot x_2 \cdot x_3\rangle \quad \text{(diss. } \cdot) \tag{17}$$

$$= \langle x_1 \cdot (x_2 \cdot x_3), x_1 \cdot (x_2 \cdot y_3 + y_2 \cdot x_3) + y_1 \cdot (x_2 \cdot x_3)\rangle \quad \text{(ass. } \cdot) \tag{18}$$

$$= \langle x_1, y_1\rangle \otimes \langle x_2 \cdot x_3, x_2 \cdot y_3 + y_2 \cdot x_3\rangle \quad \text{(def. } \otimes) \tag{19}$$

$$= \langle x_1, y_1\rangle \otimes (\langle x_2, y_2\rangle \otimes \langle x_3, y_3\rangle) \quad \text{(def. } \otimes) \tag{20}$$

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

$$\langle x_1, y_1\rangle \otimes \langle 1, 0\rangle \overset{!}{=} \langle x_1, y_1\rangle \tag{21}$$

$$\langle x_1, y_1\rangle \otimes \langle 1, 0\rangle = \langle x_1 \cdot 1, x_1 \cdot 0 + y_1 \cdot 1\rangle \quad \text{(def. } \otimes) \tag{22}$$

$$= \langle x_1, y_1\rangle \tag{23}$$

$$= \langle 1 \cdot x_1, 1 \cdot y_1 + 0 \cdot x_1\rangle \tag{24}$$

$$= \langle 1, 0\rangle \otimes \langle x_1, y_1\rangle \quad \text{(def. } \otimes) \tag{25}$$

**Axiom 3**

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

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

$$\langle x_1, y_1\rangle \otimes (\langle x_2, y_2\rangle \oplus \langle x_3, y_3\rangle) = \langle x_1, y_1\rangle \otimes \langle x_2 + x_3, y_2 + y_3\rangle \quad \text{(def. } \oplus) \tag{27}$$

$$= \langle x_1 \cdot (x_2 + x_3), x_1 \cdot (x_2 + x_3) + y_1 \cdot (y_2 + y_3)\rangle \quad \text{(def. } \otimes) \tag{28}$$

$$= \langle (x_1 \cdot x_2) + (x_1 \cdot x_3), (x_1 \cdot y_2) + (y_1 \cdot x_2) + (x_1 \cdot y_3) + (y_1 \cdot x_3)\rangle \quad \text{(diss. } \cdot) \tag{29}$$

$$= \langle x_1 \cdot x_2, (x_1 \cdot y_2) + (y_1 \cdot x_2)\rangle \oplus \langle x_1 \cdot x_3, (x_1 \cdot y_3) + (y_1 \cdot x_3)\rangle \quad \text{(def. } \oplus) \tag{30}$$

$$= (\langle x_1, y_1\rangle \otimes \langle x_2, y_2\rangle) \oplus (\langle x_1, y_1\rangle \otimes \langle x_3, y_3\rangle) \quad \text{(def. } \otimes) \tag{31}$$

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

$$(\langle x_2, y_2\rangle \oplus \langle x_3, y_3\rangle) \otimes \langle x_1, y_1\rangle = (\langle x_2 + x_3, y_2 + y_3\rangle) \otimes \langle x_1, y_1\rangle \quad \text{(def. } \oplus) \tag{33}$$

$$= \langle (x_2 + x_3) \cdot x_1, (x_2 + x_3) \cdot y_1 + (y_2 + y_3) \cdot x_1\rangle \quad \text{(def. } \otimes) \tag{34}$$

$$= \langle (x_2 \cdot x_1) \cdot (x_3 \cdot x_1), (x_2 \cdot y_1) + (x_3 \cdot y_1) + (y_2 \cdot x_1) + (y_3 \cdot x_1)\rangle \quad \text{(diss. } \cdot) \tag{35}$$

$$= \langle x_2 \cdot x_1, x_2 \cdot y_1 + y_2 \cdot x_1\rangle \oplus \langle x_3 \cdot x_1, x_3 \cdot y_1 + y_3 \cdot x_1\rangle \quad \text{(def. } \oplus) \tag{36}$$

$$= (\langle x_2, y_2\rangle \otimes \langle x_1, y_1\rangle) \oplus (\langle x_3, y_3\rangle \otimes \langle x_1, y_1\rangle) \quad \text{(def. } \otimes) \tag{37}$$

**Axiom 4**

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

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**
  $w = exp(score(t_{i-1}, t_i, \mathbf{w}))$
  $\beta(\mathbf{w}, t_i) = \oplus_{t_{i-1}\in T} \langle w, -w \cdot logw\rangle \otimes \beta(\mathbf{w}, t_{i-1})$
**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:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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:

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

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

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

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

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

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

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

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

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

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

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:

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

$$\begin{aligned}
H(T_w) &= -\sum_{t\in T^N} p(t|w)\cdot\log p(t|w) && \text{(def. } H\text{)} && (75)\\
&= -\sum_{t\in T^N} \frac{exp(score_\theta(t,w))}{Z(w)}\cdot\log(\frac{exp(score_\theta(t,w))}{Z(w)}) && \text{(def. } p\text{)} && (76)\\
&= -\sum_{t\in T^N} \frac{exp(score_\theta(t,w))}{Z(w)}\cdot(score_\theta(t,w) - \log(Z(w))) && && (77)\\
&= -\sum_{t\in T^N} \frac{exp(score_\theta(t,w))\cdot(score_\theta(t,w) - \log(Z(w)))}{Z(w)} && && (78)\\
&= \sum_{t\in T^N} \frac{exp(score_\theta(t,w))\cdot(-score_\theta(t,w) + \log(Z(w)))}{Z(w)} && && (79)\\
&= \sum_{t\in T^N} \frac{exp(score_\theta(t,w))\cdot\log(Z(w))}{Z(w)} && && (80)\\
&\quad - \sum_{t\in T^N} \frac{exp(score_\theta(t,w))\cdot score_\theta(t,w)}{Z(w)} && && (81)
\end{aligned}$$

$$\begin{aligned}
&= \sum_{t\in T^N} \frac{exp(score_\theta(t,w))\cdot\log(Z(w))}{Z(w)} && && (82)\\
&- \sum_{t\in T^N} (exp(score_\theta(t,w))\cdot score_\theta(t,w))\cdot Z(w)^{-1} && && (83)\\
&= H_U(T_w)\cdot Z(w)^{-1} + \sum_{t\in T^N} \frac{exp(score_\theta(t,w))\cdot\log(Z(w))}{Z(w)} && \text{(def. } H_U\text{)} && (84)\\
&= H_U(T_w)\cdot Z(w)^{-1} + \frac{\log(Z(w))}{Z(w)}\cdot\sum_{t\in T^N} exp(score_\theta(t,w)) && && (85)\\
&= H_U(T_w)\cdot Z(w)^{-1} + \frac{\log(Z(w))}{Z(w)}\cdot Z(w) && \text{(def. } Z(w)\text{)} && (86)\\
&= Z(w)^{-1} H_U(T_w) + \log Z(w) && && (87)
\end{aligned}$$

## d)

## d)

We want to prove that $H(T_w)$ can be calculated in $O(|T|^2 \cdot N)$ time.

From c), we can see that $H(T_w)$ is equal to $Z(w)^{-1}H_U(T_w) + \log Z(w)$. We can calculate $Z(w)$ in $O(|T|^2 \cdot N)$ time, since we have seen in the lecture that we can calculate the partition function in $O(|T|^2 \cdot N)$ time by using the forward/backward algorithm. Thus $\log Z(w)$ can also be calculated in $O(|T|^2 \cdot N)$ time. All that remains is to calculate $H_U(T_w)$, which can be done in $O(|T|^2 \cdot N)$ time, since we have shown in b) that the unnormalized entropy can be calculated in $O(|T|^2 \cdot N)$ time. This can be done by using the forward/backward algorithm over the expectaion semiring by lifting the arc weights to $\langle w, -w \cdot \log(w) \rangle$. The total time complexity is thus $O(|T|^2 \cdot N)$.

Further the derivative can be calculate in the same time complexity, since we have seen in the lecture that the derivative of a function can be calculated in the same time complexity as the function itself. Using a forward pass and a backward pass, we can calculate the derivative of $H(T_w)$ in $O(|T|^2 \cdot N)$ time.

# Question 2: Decoding a CRF with Dijkstra's Algorithm

## a)

We want to prove the following statement: "The score for the first complete tagging, popped from the priority queue is the score for the best part-of-speech tagging under the CRF."

**Prove via induction:**

Our induction hypothesis is that the first tagging of length $i$ that is popped will always be the best part-of-speech tagging of that length under the CRF.

**Base case $i = 1$:**

The base case follows trivially. All possible taggings of length one are pushed on the priority queue when $\langle \langle 0, BOT \rangle, \mathbb{1} \rangle$ is first popped. This means the first element in the priority queue will be the best part-of-speech tagging of length 1, which will be popped in the next step.

**Induction step $i \to i + 1$:**

For this case, we make a case distinction. If the best part-of-speech tagging of length $i + 1$ is the extension of the best part-of-speech tagging of length $i$ by one tag the correctness of the IH follows in the same way as in the base case. Otherwise, we use the insight that for any $x$ we have $x \geq x + y$ for any $y$ as both have to be in $\mathcal{R}_{\leq 0}$. Now let $T_{1:i+1}$ be the best part-of-speech tagging of length $i + 1$ with the previous insight it has to follow that the

score for every tagging $T_{1:k}, k \leq i+1$ is better than all other possible taggings of length $i+1$. Further at least $\langle\langle 1, t_1\rangle, s\rangle$ has to have been pushed to the queue. From this, follows that no tagging of length $i+1$ will be popped until $\langle\langle i+1, t_{i+1}\rangle, s\rangle$ has been pushed. From here the next tagging of length $i+1$ to be popped will be the tagging $T_{1:i+1}$ as it is per our assumption the best part-of-speech tagging of length $i+1$.

Thus our IH holds for all $i$ from which we can follow the statement by setting $i = N$.

## b)

Assuming we run Viterbi in a forward manner (not as presented in the lecture as then the entries would be different).

### i)

We first prove that the shape of the computed $\gamma$ is the same. Let $N$ be the length of $w$ In Viterbi, we compute for each length $i \in [1, ..., N]$ one entry per tag. This gives a $\gamma \in \mathcal{R}^{N \times |\mathcal{T}|}$ The $\gamma$ induced by Dijkstra has the same shape (neglecting the initial entry for $\langle 0, BOT\rangle$). This is very easy to see in Dijkstra we push each $\langle i, t\rangle$ pair is popped exactly once. This holds as we only push a pair if it has not yet been popped and if the pair is already queued and it is pushed again only the one with the higher score will remain in the queue. Further, each pair is pushed at least once as the algorithm, for each pair $\langle i, t\rangle$, pushes all unpopped pairs $\langle i+1, t'\rangle, t' \in T$. Thus once the algorithm finishes we will have an entry in $\gamma$ for all pairs $\langle i, t\rangle$ with $i \in [1, ..., N], t \in \mathcal{T}$ and one additional entry for $\langle 0, BOT\rangle$ which we can ignore. This gives us the same shape of $\gamma$ as Viterbi's algorithm.

### ii)

Now we prove that the entries of $\gamma$ induced by Viterbi's and Dijkstra's algorithms are the same. Both algorithms will save the score of the best part of part-of-speech tagging of length $i$ ending with tag $t$ at $\langle i, t\rangle$.

The proof for Viterbi's algorithm is simple we assume that $\gamma[i, t']$ corresponding to a tagging $T'_{1:i}$ is not the best part-of-speech tagging of length $i$ ending with tag $t$. Then there has to be a better part-of-speech tagging $T_{1:i}$ where $t_i = t'$. Now let j be the last index where $T'$ and $T$ differ this means that all tags after the j-th are the same. From this also follows that the score of the tagging $T_{1:j+1}$ has to be better than the score of $T'_{1:j+1}$ as the tags from this point are the same and $T$ scores higher. But then in the step where $\gamma[j+1, t_{j+1}]$ is assigned the algorithm would have chosen $T_{1:j}$ over $T'_{1:j}$ as we take the max overall best entries of length $j-1$. Thus $\gamma[i, t']$ has to be the score corresponding to the best part-of-speech tagging of length $i$ ending with tag $t$. Viterbi's algorithm bases entries for length $i+1$ of $\gamma$ on the entries for length $i$.

Resulting in a contradiction and proving that the entries of $\gamma$ are the best part of part-of-speech tagging of length $i$ ending with tag $t$ at $\langle i, t\rangle$.

Assume that $\gamma[i, t]$ corresponding to the tagging $T_{1:i}$ is not the score of the best part-of-speech tagging of length $i$ ending with tag $t$. From this follows that the first popped pair

for $\langle i, t \rangle$ was not the best part-of-speech tagging of length $i$ ending with tag $t$. Let $T'_{1:i}$ be the best part-of-speech tagging of length $i$ ending with tag $t$. Then as we argued in a) the scores of all $T'_{1:j}; \forall j, 1 \leq j \leq i$ have to be greater than the score of $T'_{1:i}$ this means that $\langle i, t \rangle$ would not be popped before all of $\langle j, t_j \rangle, j \in [1, ..., i]$ have been popped this means that $\gamma[i, t]$ would be set to the score of the tagging $T'$ and not $T$.

Resulting in a contradiction and proving that the entries of $\gamma$ are the best part of part-of-speech tagging of length $i$ ending with tag $t$ at $\langle i, t \rangle$ and that the entries of $\gamma$ induced by Viterbi's and Dijkstra's algorithms are the same.

## c)

Assuming a realistic priority queue implementation based on binary heaps. We have a total worst-case runtime of $\mathcal{O}(|\mathcal{T}|^2 \cdot |W| + (|\mathcal{T}| \cdot |W|) \cdot \log(|\mathcal{T}| \cdot |W|))$. As we have $|\mathcal{T}|$ edges for every vertex and $|\mathcal{T}| \cdot |W|$ vertices as we have a vertex for every tag for every word. Viterbi's algorithm has a runtime of $\mathcal{O}(|\mathcal{T}|^2 \cdot |W|)$ as we go for every word over every tag and over every previous tag. This means that in theory if $\log(|\mathcal{T}| \cdot |W|) < |\mathcal{T}|$ then it can make sense to use Dijkstra's algorithm however in reality this is very unlikely as Dijkstra also incurs more overhead due to more complicated data structures. Dijkstra can stop early and thus can be faster to calculate the best part-of-speech tagging using Dijkstra's algorithm compared to Viterbi.

## d)