Prof. Ryan Cotterell

Johanna Ribas: Assignment 3

jribas@student.ethz.ch, 22-944-003.

12/01/2023 - 17:01h

Question 1: Exploring the Kleene Star

a)

$$\begin{aligned} a^* &= \bigoplus_{n=0}^{\infty} a^n = a^0 \oplus \bigoplus_{n=1}^{\infty} a^n = \mathbf{1} \oplus \bigoplus_{n=1}^{\infty} a^n = \\ &= \mathbf{1} \oplus \bigoplus_{n=1}^{\infty} a \otimes a^{(n-1)} = \mathbf{1} \oplus a \otimes \bigoplus_{n=1}^{\infty} a^{(n-1)} = \\ &= \mathbf{1} \oplus a \otimes \bigoplus_{n=0}^{\infty} a^{(n)} = \mathbf{1} \oplus a \otimes a^* \end{aligned} \tag{1}$$

b)

$$\begin{aligned} &\mathrm{W}_{\mathrm{log}} = < \mathbb{R} \cup \{-\infty, \infty\}, \oplus_{log}, +, -\infty, 0 > \\ &1 \oplus (a \otimes a^*) = 1 \oplus (a + a^*) = \log\left(e^0 + e^{a+a^*}\right) \\ &\mathrm{a}^* = \log\left(1 + e^{a+a^*}\right) \\ &\mathrm{e}^{a^*} = 1 + e^{a+a^*} \\ &\mathrm{e}^{a^*}(1 - e^a) = 1 \\ &\mathrm{e}^{a^*} = \frac{1}{1 - e^a} \\ &a^* = \log\left(\frac{1}{1 - e^a}\right) \\ &\text{ Where } e^a \leq 1 \end{aligned} \tag{2}$$

c)

$$a^* = 1 \oplus a \otimes a^* = 1 \oplus < xx^*, xy^* + yx^* >$$

$\mathrm{x}^* = xx^* + 1 \rightarrow x^* = \frac{1}{1-x}$
$\mathrm{y}^* = xy^* + yx^* \rightarrow y^* = \frac{y}{(1-x)^2}$

$$\boxed{a^* = < \frac{1}{1-x}, \frac{y}{(1-x)^2} > \quad (3)}$$

d) 1. Proof that $\left\langle 2^{\Sigma^*}, \cup, \{\}\right\rangle$ is a commutative monoid.

1.1) Distributivity of $\oplus$

$A \cup (B \cup C) = A \cup \{x \mid x \in B \vee x \in C\} =$
$\{x \mid x \in A \vee x \in B \vee x \in C\} =$
$\{x \mid x \in A \vee x \in B] \cup C = (A \cup B) \cup C$

1.2) $\mathbf{0} \oplus A = A$ $\{\} \cup A = \{x \mid x \in A] = A$

1.3) Commutativity of $\oplus$ $\begin{aligned} A \cup B &= \{x \mid x \in A \vee x \in B\} = \\ &= \{x \mid x \in B \vee x \in A\} = B \cup A \end{aligned}$

2. Proof that $\left\langle 2^{\Sigma^*}, \otimes, \{\varepsilon\} \right\rangle$ is a monoid.

2.1) Distributivity of $\otimes$
$(A \otimes B) \otimes C =$
$\{a \circ b \mid a \in A, b \in B\} \otimes C =$
$\{a \cdot b \circ c \mid a \in A, b \in B, C \in C\} =$
$= A \otimes \{b \circ c \mid b \in B, C \in C\} = A \otimes (B \otimes C)$

2.2) $\mathbf{1} \otimes A = A$
$\{\varepsilon\} \otimes A = \{\varepsilon \circ a \mid a \in A\} = \{a \mid a \in A\} = A$

3. Proof that $\otimes$ distributes over $\oplus$
$A \otimes (B \oplus C) = A \otimes \{x \mid x \in B \vee x \in C\} =$
$= \{a \circ x \mid a \in A, x \in B \vee x \in C\} =$
$= \{a \circ x \mid a \in A, x \in B\} \vee \{a \circ x \mid a \in A, x \in C\}$
$= (A \otimes B) \oplus (B \otimes C)$

4. Proof that $\mathbf{0}$ is the anihilator.
$\{\} \otimes A = \{\}$

A kleene star A* for the language semiring is defined as:

$$A^* = \bigcup_{n=0}^{\infty} A^n = A^0 \cup A^1 \cup A^2 ... \tag{4}$$

As we did in lecture 7, we can find A^n inductively. First:

$$\begin{aligned} &A^0 = \mathbf{1} = \{\epsilon\} \\ &A^1 = A \\ &A^n = \left\{s \circ a \mid s \in A^{n-1}, a \in A\right\} \end{aligned} \tag{5}$$

Using what we proved in a):

$$A^* = \{\epsilon\} \cup (A \otimes A^*) = (A \otimes A^n) \tag{6}$$

Question 2: Asterating the matrix

a) We have to proof that $\mathbf{1} \oplus a = \mathbf{1}$.

For the tropical semiring:

$min(0, a) = 0, \forall a \in R_{\geq 0}$.

For the arctic semiring:

$max(0, a) = 0, \forall a \in R_{\leq 0}$.

b) The adjacency matrix M of graph G contains the weight of the paths of length 1 (weight of each transition between nodes). To start our proof by induction, the base case is $M^1 = M$. We assume that M^{n-1} encodes the sum of all paths of length exactly n-1. Because $M^n = M^{n-1} \otimes M$ and by the definition of matrix multiplication, we have that:

$$(M^n)_{i,k} = \bigoplus_{j=0}^{N-1} \left(M^{n-1}\right)_{i,j} \otimes M_{j,k} \tag{7}$$

This is the semiring sum of all the paths between i and k of length n-1 and length 1, i.e, it is the semiring-sum of all paths from the node i to node k in graph G of length exactly n.

c)

d) If the transition matrix is over a 0-closed semiring, the shortest path weight depends only on the paths of at most N-1.

The semiring-sum of paths between node i and node k of length at most N-1 is: $\bigoplus_{j=0}^{N-1} (M^n)_{i,k}$.

Combining these two results, we derive that:

$$M^* = \bigoplus_{n=0}^{\infty} (M^n) = \bigoplus_{n=0}^{N-1} (M^n).$$

e) For the algorithm we will use two equations:

$M^* = \bigoplus_{n=0}^{N-1} M^n$ and $M^n = M^{n-1} \otimes M$

We initialize matrix M^n and its kleene star M^* as identity matrices.

Then we run a loop N times such that at each iteration:

- Overwrite M^n with $M^n \otimes M$.
- Overwrite M^* with $M^* \oplus M^n$.

The resulting M^* will be our kleene star.

The runtime of the matrix multiplication is cubic with N. Because the algorithm computes matrix multiplication N times, the runtime of the algorithm will be $O(N^4)$.

f)

$$a \oplus a = (\mathbf{1} \otimes a) \oplus (\mathbf{1} \otimes a) = (\mathbf{1} \oplus \mathbf{1}) \otimes a = \mathbf{1} \otimes a = a \tag{8}$$

We use first the identity property and second the distributivity property. Third, because its a 0-closed semiring and $\mathbf{1}$ is an element of the semiring, $\mathbf{1} \oplus \mathbf{1} = \mathbf{1}$. Last, we use the identity property again.

g)

h)

i) First, we see that for any x: $\left\| \frac{x}{\|x\|_2} \right\|_2 = 1$. Second, if x is a constant, then $\|Ax\| = \|A\||x|$. For this, when we calculate the supremum of $\frac{\|Ax\|_2}{\|x\|_2}$, we can rewrite it as $\sup_{x\neq 0} \left\| A \frac{x}{\|x\|_2} \right\|_2 = \sup_{\|x\|_2=1} \|Ax\|_2$.

Next, we will use the Single Value Decomposition of the matrix $A = U\Sigma V^T$ where U and V are orthogonal and Σ is the diagonal matrix with $\sigma_1, ..., \sigma_n$ values. Then:

$$\sup_{\|x\|_2=1} \|Ax\|_2 = \sup_{\|x\|_2=1} \left\| U\Sigma V^T x \right\|_2 = \sup_{\|a\|_2=1} \|\Sigma a\|_2 = \sup \sqrt{\sum_{i=1}^{n} \sigma_i^2 a_i^2} = \sigma_{max}. \tag{9}$$

j)
$$\left\| A^* - \sum_{n=0}^{K} A^n \right\|_2 = \left\| \sum_{n=0}^{\infty} A^n - \sum_{n=0}^{K} A^n \right\|_2 = \left\| \sum_{n=K+1}^{\infty} A^n \right\|_2$$
$$\leq \sum_{n=K+1}^{\infty} \|A^n\|_2 \leq \sum_{n=K+1}^{\infty} \|A\|_2^n = \sum_{n=K+1}^{\infty} \sigma_{\max}(A)^n.$$

We can rewrite the previous result in terms of geometric series:

$$\sum_{n=K+1}^{\infty} \sigma_{\max}(A)^n = \sum_{n=0}^{\infty} \sigma_{\max}(A)^n - \sum_{n=0}^{K} \sigma_{\max}(A)^n$$

This substraction will go to 0 when K goes to ∞ if $|\sigma_{max}| < 1$ because then we can calculate the limit as:

$$\lim_{K\to\infty} \sum_{n=0}^{\infty} \sigma_{\max}(A)^n - \sum_{n=0}^{\Lambda} \sigma_{\max}(A)^n = \lim_{K\to\infty} \frac{1}{1-\sigma_{\max}(A)} - \frac{1-\sigma_{\max}(A)^K}{1-\sigma_{\max}(A)} = \lim_{K\to\infty} \frac{\sigma_{\max}(A)^K}{1-\sigma_{\max}(A)} = 0$$

k) The big-O bound of the truncation error is $O(\sigma_{max}(A)^K)$.

Question 3: Implementation of a Neural Transducer

Link to colab notebook:

https://colab.research.google.com/drive/1ymghMoIvWvvLeUs63ruBn1PTGI_Tl6E9?usp=sharing