Natural Language Processing: Assignment 3

Isaac Jefferson Lee

Email: isalee@student.ethz.ch

Question 1

Question 1 :: Part a)

We are given that:

$$a^* := \bigoplus_{n=0}^{\infty} a^{\otimes n}.$$

So it follows that:

$$a \otimes a^* = a \otimes \left(\bigoplus_{n=0}^{\infty} a^{\otimes n} \right).$$

By the distributive property:

$$= \bigoplus_{n=0}^{\infty} (a \otimes a^{\otimes n}) = \bigoplus_{n=0}^{\infty} a^{\otimes (n+1)}.$$

We can re-label our index and clearly we have that:

$$a \otimes a^* = \bigoplus_{m=1}^{\infty} a^{\otimes m}.$$

Then it follows that:

$$\mathbf{1} \oplus a \otimes a^* = \mathbf{1} \oplus \bigoplus_{m=1}^{\infty} a^{\oplus m}.$$

Note: We must have that $a^{\otimes 0} = \mathbf{1}$, (easy to prove by contradiction), so then we have:

$$\mathbf{1} \oplus \bigoplus_{m=1}^{\infty} a^{\otimes m} = \bigoplus_{m=0}^{\infty} a^{\otimes m} =: a^*.$$

QED.

Question 1 :: Part b)

So we want some $a^* \in \mathbb{R} \cup \{-\infty\}$ such that:

$$\begin{aligned}
a^* &= \mathbf{1} \oplus a \otimes a^* \\
&= 0 \oplus_{\log} (a + a^*) \\
&= \log(1 + \exp(a + a^*)) \\
&= \log(1 + \exp(a)\exp(a^*)) \\
&\implies \exp(a^*) = 1 + \exp(a)\exp(a^*) \\
&\implies \exp(a^*)(1 - \exp(a)) = 1 \\
&\implies \exp(a^*) = \frac{1}{1 - \exp(a)} \\
&\implies a^* = \log\left(\frac{1}{1 - \exp(a)}\right) \\
&= -\log(1 - \exp(a)).
\end{aligned}$$

Note: Clearly this is only defined for $a < 0$.

Question 1 :: Part c)

As before, we seek some a^* s.t:

$$a^* = \mathbf{1} \oplus a \otimes a^*.$$

Translating this to the expectation semiring, we require some $a^* := \langle a_1^*, a_2^* \rangle$ s.t:

$$\begin{aligned}
\langle a_1^*, a_2^* \rangle &= \langle 1, 0 \rangle \oplus (\langle a_1, a_2 \rangle \otimes \langle a_1^*, a_2^* \rangle) \\
&= \langle 1, 0 \rangle \oplus \langle a_1 * a_1^*, a_1 * a_2^* + a_2 * a_1^* \rangle \\
&= \langle 1 + a_1 * a_1^*, a_1 * a_2^* + a_2 * a_1^* \rangle
\end{aligned}$$

.

So clearly this first term is just the Kleene Star over the Real Semiring, so it follows that:

$$a_1^* = \sum_{n=0}^{\infty} a_1^n = \frac{1}{1 - a_1}.$$

(Assuming $|a_1| < 1$ for convergence). So now we want some a_2^* s.t:

$$a_2^* = a_1 * a_2^* + a_2 * \frac{1}{1 - a_1}.$$

Therefore we can re-arrange this to get:

$$a_2^*(1 - a_1) = \frac{a_2}{1 - a_1} \implies a_2^* = \frac{a_2}{(1 - a_1)^2}.$$

Then combining we get:

$$a^* := \langle a_1^*, a_2^* \rangle = \left\langle \frac{1}{1 - a_1}, \frac{a_2}{(1 - a_1)^2} \right\rangle.$$

Question 1 :: Part d)

$$\mathcal{W}_{\text{lang}} := \langle 2^{\Sigma^*}, \cup, \otimes, \phi, \{\epsilon\}\rangle.$$

Where $A \otimes B := \{a \circ b \;\; s.t \;\; a \in A, b \in B\}$

First we check that $\langle 2^{\Sigma^*, \cup, \phi}\rangle$ is a commutative monoid. Clearly from elementary set theory we know that union is both commutative and associative.

Also we know that $A \cup \phi = A = \phi \cup A \; \forall A \in 2^{\Sigma^*}$ so clearly ϕ is the left and right unit and so we have verified that $\langle 2^{\Sigma^*, \cup, \phi}\rangle$ is a commutative monoid.

Next we must check that $\langle 2^{\Sigma^*}, \otimes, \{\epsilon\}\rangle$ is a monoid. Note that:

$$\begin{aligned} A \otimes (B \otimes C) &= A \otimes \{b \circ c \;\; s.t \;\; b \in B, c \in C\} \\ &= \{a \circ b \circ c \;\; s.t \;\; b \in B, c \in C\} \end{aligned}$$

since string concatenation is associative. Then $(A \otimes B) \otimes C$ is equivalent by symmetry and so we get associativity. Next we see that:

$$A \otimes \{\epsilon\} := \{a \circ \epsilon \;\; s.t \;\; a \in A\} = A.$$

Also:

$$\{\epsilon\} \otimes A := \{\epsilon \circ a \;\; s.t \;\; a \in A\} = A.$$

and so clearly $\{\epsilon\}$ is the left and right unit and thus we have verified that $\langle 2^{\Sigma^*}, \oplus, \{\epsilon\}\rangle$ is a monoid.

Next we show that $\otimes$ distributes over $\oplus := \cup$. Suppose $A, B, C \in 2^{\Sigma^*}$, then:

$$\begin{aligned} (A \cup B) \otimes C &= \{\alpha \circ c \;\; s.t \;\; \alpha \in A \cup B, c \in C\} \\ &= \{a \circ c \;\; s.t \;\; a \in A, c \in C\} \cup \{b \circ c \;\; s.t \;\; b \in B, c \in C\} \\ &= (A \otimes C) \cup (B \otimes C) \end{aligned}$$

.

Which proves the first part of distributivity and then it is clear to see by symmetry that the second part of distributivity also holds.

Finally we must show that ϕ is an annihilator for $\otimes$. Suppose $A \in 2^{\Sigma^*}$ then $\phi \otimes A = \{\gamma \circ a \;\; s.t \;\; \gamma \in \phi, a \in A\} = \phi$. Since $\exists\, \gamma \in \phi$ can never be true. Similarly for $A \otimes \phi$. So we have verified all of the semiring axioms.

Now we are interested in finding a Kleene Star for $\mathcal{W}_{\text{lang}}$, i.e we want some $A^* \in 2^{\Sigma^*}$ s.t:

$$A^* = \{\epsilon\} \cup \{a \circ a' \;\; s.t \;\; a \in A, a' \in A^*\}.$$

Clearly the Kleene Closure Σ^* satisfies this. This simply follows from the recursive definition of Σ^* given by:

$$\Sigma^0 := \{\epsilon\}, \quad \Sigma^1 := \Sigma, \quad \forall i \in \mathbb{N}, \Sigma^{i+1} := \{w \circ a \;\; s.t \;\; w \in \Sigma^i, a \in \Sigma\}.$$

$$\Sigma^* := \bigcup_{i \in \mathbb{N}} \Sigma^i.$$

Question 2

Question 2 :: Part a)

Recall the Tropical Semiring: $\langle \mathbb{R}_{\geq 0}, min, +, \infty, 0 \rangle$. Suppose we have some arb. $a \in \mathbb{R}_{\geq 0}$, then $\mathbf{1} \oplus a := min(0, a) = 0 =: \mathbf{1}$, since $a \in R_{\geq 0} \implies a \geq 0$. So clearly we have the 0-closed property.

Recall the Arctic Semiring: $\langle \mathbb{R}_{\leq 0}, max, +, -\infty, 0 \rangle$. Suppose we have some arb. $b \in \mathbb{R}_{\leq 0}$, then $\mathbf{1} \oplus b := max(0, b) = 0 =: \mathbf{1}$, since $b \in \mathbb{R}_{\leq 0} \implies b \leq 0$. So clearly we have the 0-closed property.

Question 2 :: Part b)

We proceed by induction on n. When $n = 1$ the result is trivial. Now assume true for $n = m \in \mathbb{N}$. Then when $n = m + 1$ we have $M^{m+1} = M \otimes M^m$, so:

$$(M^{m+1})_{ij} = \bigoplus_{k=1}^{N} M_{ik} \otimes (M^m)_{kj} = \bigoplus_{k=1}^{N} w_{ik} \otimes (M^m)_{kj}.$$

So from our induction hypothesis we know that $(M^m)_{kj}$ is the sum of paths of length exactly m, starting at k and ending at j. Clearly $M_{ik} \otimes (M^m)_{kj}$ represents all possible paths with a single edge from i to k and then the next edges starting at k and ending at j. So then since we are summing over all possible intermediate k, it follows that we consider all possible paths of length $m+1$ and therefore we have that $(M^{m+1})_{ij}$ is the sum of all possible paths of length $m+1$ starting at node i and ending at j and we are done.

QED.

Question 2 :: Part c)

Main Idea: Paths of length $> N - 1$ must contain cycles.

Suppose we have some path from i to j of length greater than $N - 1$. It is clear that this path will contain atleast one cycle, since we must visit atleast one node more than once. So we can factor the path weight as:

$$w(\pi) = \underbrace{\left(\bigotimes_{k=1}^{p} w_k \right)}_{:= w_\alpha(\pi)} \otimes \underbrace{\left(\bigotimes_{k=1}^{q} w'_k \right)}_{:= w_\beta(\pi)}.$$

where $p + q = M$ and w'_k are weights from the edges of π that we remove to make π a path from i to j with no-cycles. (Note that for large paths with many cycles there could be many different choices of such weights but this choice is without loss of generality).

So since the non-removed edges contain no cycles, it follows that $p \leq N - 1$. So using our notation, $w(\pi) = w_\alpha(\pi) \otimes w_\beta(\pi)$. It follows by definition of $Z(i, j)$ that since $w_\alpha(\pi)$ is a valid path from i to j then both $w_\alpha(\pi)$ and $w_\alpha(\pi) \otimes w_\beta(\pi)$ will be part of the semiring-plus sum given by (2). So in the sum we will have:

$$w_\alpha(\pi) \oplus (w_\alpha(\pi) \otimes w_\beta(\pi)).$$

By distributivity:

$$= w_\alpha(\pi) \otimes (\mathbf{1} \oplus w_\beta(\pi)).$$

Then by the 0-closed property:

$$= w_\alpha(\pi) \otimes (\mathbf{1}) = w_\alpha(\pi).$$

and we can clearly do this for any path of length $> N-1$ and so we have shown that the sum only depends on paths of length $\leq N-1$. QED.

Question 2 :: Part d)

Recall: M^* must satisify:

$$M^* = \mathbf{1} \oplus M \oplus M^*.$$

So defining $M^* := \bigoplus_{n=0}^{N-1} M^n$ we must verify the above.

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

By distributivity:

$$= \bigoplus_{n=0}^{N-1} M^{n+1}.$$

Re-labelling we get:

$$= \bigoplus_{r=1}^{N} M^r.$$

$$\implies \mathbf{1} \oplus M \otimes M^* = \bigoplus_{r=0}^{N} M^r.$$

since we must have that $M^0 = \mathbf{1}$. Then using part b) we know that M^N encodes the paths of length N and by part c) we showed that the sum is independent of paths longer than $N-1$ and so we see that M^N does not contribute to the sum and we get:

$$\bigoplus_{r=0}^{N} M^r = \bigoplus_{r=0}^{N-1} M^r =: M^*.$$

and we are done. QED.

Question 2 :: Part e)

The for loop gives:

$$\mathbf{0} \oplus (\mathbf{1} \otimes M) \oplus (\mathbf{1} \otimes M \otimes M) \oplus ... \oplus (\mathbf{1} \otimes M \otimes ... \otimes M).$$

So clearly by definition of M^n we are computing $\bigoplus_{n=0}^{N-1}$ as required. If we assume that matrix multiplication has runtime complexity $O(N^3)$ using our naive algorithm we do $N-1$ matrix multiplications therefore we have a worst case runtime complexity:

$$O(N * N^3) = O(N^4).$$

Algorithm 1 Naive M^*

```
def m_star(M)
M* ← 1
prod ← 1
for n ∈ {1, ..., N − 1} do
    prod ← prod ⊗ M
    M* ← M* ⊕ prod
end for
     return M*
```

Question 2 :: Part f)

$$0\text{-closed} \implies \mathbf{1} \oplus \mathbf{1} = \mathbf{1} \implies a \otimes (\mathbf{1} \oplus \mathbf{1}) = a \otimes \mathbf{1} = a.$$

And by distributivity:

$$a \otimes (\mathbf{1} \oplus \mathbf{1}) = a \oplus a \implies a \oplus a = a.$$

QED.

Question 2 :: Part g)

By the binomial expansion we get:

$$(I \oplus M)^K = \bigoplus_{n=0}^{K} \binom{K}{n} I^{K-n} M^n.$$

By the identity property

$$= \bigoplus_{n=0}^{K} \binom{K}{n} M^n.$$

Clearly by idempotency, for any natural number, $\alpha \in \mathbb{N}$:

$$\alpha M^n := \underbrace{M^n \oplus M^n \oplus \ldots \oplus M^n}_{\alpha \text{ times}} = M^n.$$

$$\implies \bigoplus_{n=0}^{K} \binom{K}{n} M^n = \bigoplus_{n=0}^{K} M^n$$

and we are done. ED.

Question 2 :: Part h)

Recall that $M^n := \underbrace{M \otimes M \otimes \ldots \otimes M}_{n \text{ times}}$ so in order for:

$$M^n = \bigotimes_{k=0}^{\lfloor \log_2(n) \rfloor} M^{\alpha_k 2^k}.$$

to hold, we need to choose α_k such that:

$$\sum_{k=0}^{\lfloor \log_2(n) \rfloor} \alpha_k 2^k = n.$$

clearly we can represent n in binary and therefore such α_k exist. (α_k is given by the k^{th} digit of the binary representation of n).

$$(3) + (g) \implies M^* = (I \oplus M)^{N-1}.$$

$$(4) \implies (I \oplus M)^{N-1} = \bigotimes_{k=0}^{\lfloor \log_2(N-1) \rfloor} (I \oplus M)^{\alpha_k 2^k}.$$

Algorithm 2 Faster M^*

```
def m_star_faster(M)
α ← int_to_binary(N − 1)
M* ← 1
prod ← 1
for k ∈ {0, ..., ⌊log2(N − 1)⌋} do
    if α_k == 1 then
        prod ← prod ⊗ (I ⊕ M)
        M* ← M* ⊗ prod
    end if
end for
    return M*
```

We do two matrix multiplications, of the order $\log(N)$ times, so we have overall runtime complexity $O(\log(N)N^3)$.

Question 2 :: Part i)

Note: If $A \in \mathbb{R}^{n \times m}$ is a matrix, then the singular values of A are given by the square roots of the eigenvalues of $A^T A$.

$$||A||_2 := \sup_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{||A\boldsymbol{x}||_2}{||\boldsymbol{x}||_2} = \sup_{||\boldsymbol{x}||_2 = 1} ||A\boldsymbol{x}||_2.$$

By singular value composition we have that:

$$\sup_{||x||_2 = 1} ||A\boldsymbol{x}||_2 = \sup_{||\boldsymbol{x}||_2 = 1} ||U\Sigma V^T \boldsymbol{x}||_2 = \sup_{||\boldsymbol{x}||_2 = 1} ||\Sigma V^T \boldsymbol{x}||_2.$$

Where the last equality follows from the fact that U is unitary.

Define $\boldsymbol{y} := V^T \boldsymbol{x}$ and then it follows that $||\boldsymbol{y}||_2 = 1$ since V is also unitary. So we have that:

$$\sup_{||\boldsymbol{x}||_2 = 1} ||\Sigma V^T \boldsymbol{x}||_2 = \sup_{||\boldsymbol{y}||_2 = 1} ||\Sigma \boldsymbol{y}||_2.$$

Now since we know that $\Sigma_{ij} = \sigma_i$ whenever $i = j$ and 0 otherwise, then and we also defined Σ s.t $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_m$, then we see that the max is attained when $\boldsymbol{y} = \boldsymbol{e}_1$. This is easy to see from the fact that:

$$\sqrt{\sum_{i=1}^{m} (\sigma_i y_i)^2} \leq \sqrt{\sum_{i=1}^{m} \sigma_i^2 y_i^2} \leq \sqrt{\sum_{i=1}^{m} \sigma_{\max}^2 y_i^2} \leq \sigma_{\max} \sqrt{\sum_{i=1}^{m} y_i^2} = \sigma_{\max}.$$

and we see that $||A||_2 = \sigma_1 = \sigma_{\max}(A)$

QED.

Question 2 :: Part j)

$A^* = \sum_{n=0}^{\infty} A^n$ so it follows that:

$$\begin{aligned}
||A^* - \sum_{n=0}^{k} A^n||_2 &= ||\sum_{n=0}^{\infty} A^n - \sum_{n=0}^{k} A^n||_2 \\
&= ||\sum_{n=k+1}^{\infty} A^n + \sum_{n=0}^{k} A^n - \sum_{n=0}^{k} A^n||_2 \\
&= ||\sum_{n=k+1}^{\infty} A^n||_2 \\
&= ||\lim_{m\to\infty} \sum_{n=k+1}^{m} A^n||_2 \\
&= \lim_{m\to\infty} ||\sum_{n=k+1}^{m} A^n||_2
\end{aligned}$$

.

By the triangle inequality:

$$\leq \lim_{m\to\infty} \sum_{n=k+1}^{m} ||A^n||_2.$$

By Cauchy-Schwartz for the operator norm:

$$\leq \lim_{m\to\infty} \sum_{n=k+1}^{m} ||A||_2^n = \lim_{m\to\infty} \sum_{n=k+1}^{m} \sigma_{\max}(A)^n.$$

So for any fixed k, we need this error to converge, so by the root test for infinite series, if we define $a_n := \sigma_{\max}(A)^n$, then:

$$L := \lim_{n\to\infty} |a_n|^{\frac{1}{n}} = \lim_{n\to\infty} \sigma_{\max}(A) = \sigma_{\max}(A).$$

So it follow that for convergence we require $\sigma_{\max}(A) < 1$. Then we get:

$$\begin{aligned}
\sum_{n=k+1}^{\infty} \sigma_{max}^n &= \sum_{n=0}^{\infty} \sigma_{max}^n - \sum_{n=0}^{k} \sigma_{max}^n \\
&= \frac{1}{1-\sigma_{max}} - \frac{1-\sigma_{max}^{k+1}}{1-\sigma_{max}} = \frac{\sigma_{max}^{k+1}}{1-\sigma_{max}}
\end{aligned}$$

.

Question 2 :: Part k)

So from part j) we have:

$$||A^* - \sum_{n=0}^{k} A^n||_2 = \frac{\sigma_{\max}(A)^{k+1}}{1-\sigma_{\max}(A)}.$$

Recall that $f(k) = O(\phi(k))$ as $k \to \infty$ means $|f(k)/\phi(k)|$ is bounded for sufficiently large k. So clearly $\phi(k) = \sigma_{\max}(A)^k$ and we have

$$\left|\frac{f(k)}{\phi(k)}\right| = \frac{\sigma_{\max}(A)}{1-\sigma_{\max}(A)}.$$

which is certainly bounded for all k. So our big-O bound is $O(\sigma_{\max}(A)^k)$ and we see that our error gets exponentially small as k gets large so it is a good approximation.

Question 3

See uploaded .ipynb.