This list of exercises simulates an exam for the course *Natuurlijke Taalmodellen en Interfaces*. Answer the questions in the spaces provided. If you run out of space for an answer, continue on the back of the page.
Mobile phones, tablets, computers, e-readers, and other electronic equipments are not allowed. They must be switched off and stored away. Basic calculators (not scientific ones) are allowed, but not required, neither necessary.

Contents

Points

Question:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Total
Points:	1	1	3	1	2	1	4	3	3	2	1	2	2	3	1	1	2	4	6	2	45

1 Random variables and rules of probabilities

1. (1 point) Let X be a random variable whose sample space is the English vocabulary Σ and whose mapping to $\mathbb{R}$ is realised by an arbitrary *enumeration* of Σ. Given the partial definitions of X below, mark those that are definitely **invalid**?

✗ $$X(\omega) = \begin{cases} 1 & \textbf{if } \omega = \{\textbf{the}\} \\ 2 & \textbf{if } \omega = \{\textbf{the}\} \\ 3 & \textbf{if } \omega = \{\textbf{cat}\} \\ 4 & \textbf{if } \omega = \{\textbf{dog}\} \\ \ldots \end{cases}$$

✗ $$X(\omega) = \begin{cases} 1 & \textbf{if } \omega = \{\textbf{the}\} \\ 1 & \textbf{if } \omega = \{\textbf{a}\} \\ 2 & \textbf{if } \omega = \{\textbf{cat}\} \\ 3 & \textbf{if } \omega = \{\textbf{dog}\} \\ \ldots \end{cases}$$

✗ $$X(\omega) = \begin{cases} 1 & \textbf{if } \omega = \{\textbf{the}, \textbf{a}\} \\ 2 & \textbf{if } \omega = \{\textbf{cat}\} \\ 3 & \textbf{if } \omega = \{\textbf{dog}\} \\ \ldots \end{cases}$$

○ $$X(\omega) = \begin{cases} 1 & \text{if } \omega = \{\text{the}\} \\ 2 & \text{if } \omega = \{\text{a}\} \\ 3 & \text{if } \omega = \{\text{cat}\} \\ 4 & \text{if } \omega = \{\text{dog}\} \\ \ldots \end{cases}$$

Solution: A valid enumeration should not contain duplicates (which excludes the top-left example) and should map one-to-one (which excludes the bottom-left and the top-right example).

2. (1 point) Number the identities on the right according to the concepts on the left.

1. Chain rule	$(\underline{\;\;\mathbf{2}\;\;})$ $P_{A\mid B}(a\mid b) = \frac{P_{AB}(a,b)}{P_B(b)}$
2. Conditional probability	$(\underline{\;\;\mathbf{4}\;\;})$ $P_A(a) = \sum_{b \in \mathcal{B}} P_{AB}(a,b)$
3. Bayes rule	$(\underline{\;\;\mathbf{1}\;\;})$ $P_{AB}(a,b) = P_B(b)P_{A\mid B}(a\mid b)$
4. Marginal probability	$(\underline{\;\;\mathbf{3}\;\;})$ $P_{B\mid A}(b\mid a) = \frac{P_B(b)P_{A\mid B}(a\mid b)}{P_A(a)}$

2 Categorical distributions

3. Let X be a Categorical random variable:

$$X \sim \mathrm{Cat}(\theta_1, \ldots, \theta_v)$$

(a) (½ point) What is the support $\mathcal{X}$ of the random variable?

> **Solution:** The set $\mathcal{X} = \{1, \ldots, v\}$

(b) (½ point) What is the value of $P_X(x)$?

> **Solution:** $P_X(x) = \theta_x$

(c) (1 point) What conditions apply to valid parameters $\langle\theta_1, \ldots, \theta_v\rangle$?

> **Solution:** Every parameter must be a probability value, thus $0 < \theta_x < 1$, and together they must sum to 1, i.e. $\sum_{x=1}^{v} \theta_x = 1$

(d) (1 point) Given a data set of n i.i.d. observations, what is the maximum likelihood estimate of θ_x?

> **Solution:** The relative frequency of the class x in the dataset. That is,
>
> $$\theta_x = \frac{\sum_{i=1}^{n}[x_i = x]}{n} = \frac{\mathrm{count}_X(x)}{n}$$
>
> where $[\cdot]$ is the Iverson bracket.

Total for Question 3: 3

4. (1 point) Select, out of the list below, vector(s) that constitute(s) **valid** categorical parameters for a categorical random variable that may take on one out of 7 classes.

✗ $\langle 0.1, 0.1, 0.1, 0.1, 0.1, 0.2, 0.3\rangle$
○ $\langle 0.1, 0.1, 0.1, 0.1, 0.1, 0.5\rangle$
○ $\langle 0.2, 0.2, 0.1, 0.1, 0.1, 0.2, 0.2\rangle$

> **Solution:** A valid parameter vector should contain 7 probability values (because we have 7 classes), they must be real numbers between 0 and 1, and together they must sum to 1.

3 Markov models

5. Consider the probability of a sentence as given by the following factorisation

$$\begin{aligned} P_S(x_1^n) &= P_N(n) P_{S|N}(x_1^n|n) \\ &= P_N(n) \prod_{i=1}^{n} P_{X|H}(x_i|x_{<i}) \end{aligned}$$

where S is a random sentence, N a random length, X a random word, and H a random history.

(a) (½ point) Select appropriate descriptions for x_1^n

✗ **an outcome of S**

✗ **a sequence of n random words**

○ n outcomes of S

(b) (½ point) Select appropriate descriptions for n

✗ **a random length**

○ a random noun

✗ **the length of the outcome of S**

(c) (½ point) Select appropriate descriptions for x_i

✗ **a random word**

✗ **the ith element of the outcome of S**

○ the ith random sequence

(d) (½ point) Select appropriate descriptions for $x_{<i}$

○ a word if $i = 2$

✗ **a random sequence**

✗ **the ith random history**

Total for Question 5: 2

6. (1 point) Let x_1^n be the outcome of a random sentence S, and let $P_{S|N}(x_1^n|n)$ denote its probability value (given length n) under a **unigram** language model. Write down the expression that corresponds to this probability value.

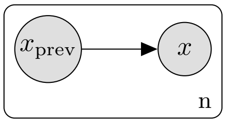
Solution: The joint probability of x_1^n given n can be expressed via the chain rule

$$P_{S|N}(x_1^n|n) = \prod_{i=1}^{n} P_{X|H}(x_i|x_{<i})$$

then under a 0-order Markov assumption we get a *unigram* language model, which factorises as

$$P_{S|N}(x_1^n|n) = \prod_{i=1}^{n} P_X(x_i)$$

7. Answer questions about the graphical model below, where X is a random variable over exactly v English words.



(a) (½ point) Which language model (LM) is this?

A. unigram LM **B. bigram LM** C. hidden Markov LM

(b) (½ point) How many conditional probability distributions (cpds) are there in the model (ignore the *length* distribution)?

A. one B. two C. n **D.** v

(c) (½ point) Is $P_{X|X_{\text{prev}}=x_{\text{prev}}}$ a tabular cpd or an inferred distribution?

A. tabular B. inferred

(d) (½ point) Is $P_{S|N=n}$ a tabular cpd or an inferred distribution?

A. tabular **B. inferred**

(e) (1 point) Write down the expression of the probability value $P_S(x_1^n)$ (you may assume appropriate padding exists).

Solution: Assuming x_0 corresponds to some BoS symbol, we have

$$P_S(x_1^n) = P_N(n) \prod_{i=1}^{n} P_{X|X_{\text{prev}}}(x_i|x_{i-1})$$

(f) (½ point) Assume that the probability value $P_{X|X_{\text{prev}}}(x|x_{\text{prev}})$ can be assessed in constant time. Express the complexity of computing $P_{S|n}(x_1^n|n)$ as a function of sentence length (use *big-O-notation*).

Solution: The joint probability contains one conditional probability value per word, thus the complexity is $O(n)$

(g) (½ point) Suppose we have exactly v words in the vocabulary, and we use a Categorical distribution for each cpd in the model. What is the representation cost of this model (use *big-O-notation*)?

Solution: Each cpd costs $O(v)$ and we have v cpds in the model, thus the representation cost is $O(v^2)$.

Total for Question 7: 4

8. Consider the following unigram language model, where EoS is a special symbol deterministically added to the end of every sentence, and answer the questions below. In this exercise you are expected to pad sentences with a BoS token, which **is not** modelled, and an EoS token, which **is** modelled.

X	$\text{Cat}(x\|\boldsymbol{\theta})$
a	θ_{a}
b	θ_{b}
c	θ_{c}
d	θ_{d}
EoS	θ_{EoS}

(a) (½ point) What is the probability of the sentence a b c a d given its length?

Solution:

$$\begin{aligned} P_{S|N}(\langle \text{a}, \text{b}, \text{c}, \text{a}, \text{d}, \text{EoS}\rangle | n) &= \theta_{\text{a}} \times \theta_{\text{b}} \times \theta_{\text{c}} \times \theta_{\text{a}} \times \theta_{\text{d}} \times \theta_{\text{EoS}} \\ &= \theta_{\text{a}}^2 \times \theta_{\text{b}} \times \theta_{\text{c}} \times \theta_{\text{d}} \times \theta_{\text{EoS}} \end{aligned}$$

(b) (½ point) What is the probability of the sentence a b b d c a a f?

Solution: Without any smoothing this sentence has probability 0 because it contains a word which is not in the support of the unigram distribution, i.e. *f*.

(c) (1 point) What is the role of smoothing?

Solution: Smoothing allows us to reserve probability mass to unseen events, which ultimately allows us to assign non-zero probabilities to every sentence that may ever occur.

(d) (1 point) Answer true (T) or false (F).

i. **T** The sentence a a b c has the same probability as the sentence a b a c.

ii. **F** The unigram language model is sensitive to word order.

iii. **T** A smoothed unigram language model has infinite support.

iv. **T** Without smoothing, and without taking padding into account, the support of the unigram language model above is the set of strings in $\{\text{a}, \text{b}, \text{c}, \text{d}\}^*$.

Total for Question 8: 3

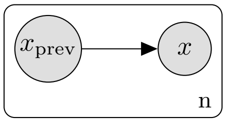
9. Consider the generative story below

$$
\begin{aligned}
N &\sim P_N \\
X_i | X_{i-1} = x_{i-1} &\sim \mathrm{Cat}(\theta_1^{(x_{i-1})}, \ldots, \theta_v^{(x_{i-1})}) \qquad\qquad \text{for } i = 1, \ldots, n
\end{aligned}
$$

and assume that every distribution in the model gives support to $\{\text{a}, \text{b}, \text{c}, \text{d}, \text{Unk}\}$, where Unk is a special token to which we map all unseen words, in addition to a EoS padding symbol that occurs always at the end of strings.

(a) (1 point) Draw the graphical model using plate notation.

Solution:



(b) Pad the sentence a b c a b and answer the questions below.

i. (1 point) List its bigrams and their counts.

Solution: Padded sentence: $\langle \text{BoS}, \text{a}, \text{b}, \text{c}, \text{a}, \text{b}, \text{EoS} \rangle$.

Bigram	Count
BoS a	1
a b	2
b c	1
c a	1
b EoS	1

ii. (1 point) What is the probability of the sentence given its length? Express probability values in terms of the parameters shown in the generative story.

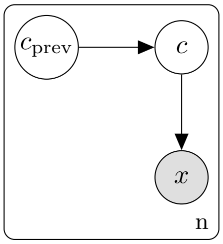
Solution:

$$
\begin{aligned}
P_{S|N}(x_1^n|n) &= \theta_{\text{a}}^{(\text{BoS})} \times \theta_{\text{b}}^{(\text{a})} \times \theta_{\text{c}}^{(\text{b})} \times \theta_{\text{a}}^{(\text{c})} \times \theta_{\text{b}}^{(\text{a})} \times \theta_{\text{EoS}}^{(\text{b})} \\
&= \theta_{\text{a}}^{(\text{BoS})} \times \left(\theta_{\text{b}}^{(\text{a})}\right)^2 \times \theta_{\text{c}}^{(\text{b})} \times \theta_{\text{a}}^{(\text{c})} \times \theta_{\text{EoS}}^{(\text{b})}
\end{aligned}
$$

Total for Question 9: 3

4 Hidden Markov models

The hidden Markov model (HMM) extends the Markov model with word categories. The graphical model below specifies the conditional independence assumptions of the HMM, where



X is a random word from a vocabulary of v words and C is a random word category (or tag) from a vocabulary of t tags. There are two types of cpds in the HMM. *Transition distributions* used to generate a tag given the tag of the previous word:

$$C|C_{\text{prev}} = c_{\text{prev}} \sim \text{Cat}(\lambda_1^{(c_{\text{prev}})}, \dots \lambda_t^{(c_{\text{prev}})})$$

And *emission distributions* used to generate a word given its tag:

$$X|C = c \sim \text{Cat}(\theta_1^{(c)}, \dots, \theta_v^{(c)})$$

The joint probability for a sentence x_1^n and tag-sequence c_1^n given length $N = n$ factorises

$$\begin{aligned} P_{X_1^n C_1^n|N}(x_1^n, c_1^n|n) &= P_{X_1^n C_1^n|N}(x_1^n, c_1^n|n) \\ &= \prod_{i=1}^{n} P_{C|C_{\text{prev}}}(c_i|c_{i-1}) P_{X|C}(x_i|c_i) \end{aligned}$$

in terms of transition and emission probabilities.

Assessing the probability of a sentence, regardless of tag sequence, requires marginalisation

$$\begin{aligned} P_{X_1^n|N}(x_1^n|n) &= \sum_{c_1=1}^{t} \cdots \sum_{c_n=1}^{t} P_{X_1^n C_1^n|N}(x_1^n, c_1^n|n) \\ &= \prod_{i=1}^{n} \sum_{c_{i-1=1}}^{t} \sum_{c_{i=1}}^{t} P_{C|C_{\text{prev}}}(c_i|c_{i-1}) P_{X|C}(x_i|c_i) \end{aligned}$$

10. (2 points) Express the joint probability for a sentence x_1^n and tag-sequence c_1^n given their length as a function of transition and emission parameters.

Solution:

$$P_{X_1^n C_1^n|N}(x_1^n, c_1^n|n) = \prod_{i=1}^{n} \lambda_{c_i}^{(c_{i-1})} \times \theta_{x_i}^{(c_i)}$$

11. (1 point) The HMM has $\underline{\quad t \quad}$ transition distributions, each of which has $\underline{\quad t \quad}$ parameters, it also contains $\underline{\quad t \quad}$ emission distributions, each of which has $\underline{\quad v \quad}$ parameters, and therefore, the total representation cost of the HMM (in big-O-notation) is $\underline{O(t^2 + t \times v)}$.

12. (2 points) Write the generative story of the HMM (you may assume appropriate padding exists).

Solution: Assuming c_0 corresponds to a BoS symbol, and for $i = 1, \ldots, n$

$$\begin{aligned} N &\sim P_N \\ C_i|C_{i-1} = c_{i-1} &\sim \mathrm{Cat}(\lambda_1^{(c_{i-1})}, \ldots, \lambda_t^{(c_{i-1})}) \\ X_i|C_i = c_i &\sim \mathrm{Cat}(\theta_1^{(c_i)}, \ldots, \theta_v^{(c_i)}) \end{aligned}$$

13. (2 points) Consider the following transition and emission distributions.

	$C=1$	$C=2$	$C=3$
$C_{\text{prev}}=0$	$\lambda_1^{(0)}$	$\lambda_2^{(0)}$	$\lambda_3^{(0)}$
$C_{\text{prev}}=1$	$\lambda_1^{(1)}$	$\lambda_2^{(1)}$	$\lambda_3^{(1)}$
$C_{\text{prev}}=2$	$\lambda_1^{(2)}$	$\lambda_2^{(2)}$	$\lambda_3^{(2)}$
$C_{\text{prev}}=3$	$\lambda_1^{(3)}$	$\lambda_2^{(3)}$	$\lambda_3^{(3)}$

	$X=1$	$X=2$	$X=3$	$\dots$	$X=v$
$C=1$	$\theta_1^{(1)}$	$\theta_2^{(1)}$	$\theta_3^{(1)}$	$\dots$	$\theta_v^{(1)}$
$C=2$	$\theta_1^{(2)}$	$\theta_2^{(2)}$	$\theta_3^{(2)}$	$\dots$	$\theta_v^{(2)}$
$C=3$	$\theta_1^{(3)}$	$\theta_2^{(3)}$	$\theta_3^{(3)}$	$\dots$	$\theta_v^{(3)}$

Transition distributions (left) and emission distributions (right)

We can use an HMM model defined with these cpds to find the best possible way to tag an input sentence $\langle x_1, x_2, x_3\rangle$. The table below shows 3 cells used to compute the Viterbi recursion $\alpha(i,j)$. What is the value of the Viterbi entry $\alpha(i=2, j=1)$?

	$i=1$	$i=2$	$i=3$
$C=1$	$\lambda_1^{(0)}\theta_{x_1}^{(1)}$	?	
$C=2$	$\lambda_2^{(0)}\theta_{x_1}^{(2)}$		
$C=3$	$\lambda_3^{(0)}\theta_{x_1}^{(3)}$		

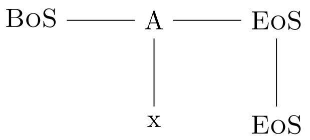
Viterbi table $\alpha(i,j)$: assume $j=0$ to correspond to the BoS tag.

Solution: We have to consider all possible ways in which we can arrive at $\langle C_2 = 1, X_2 = x_2\rangle$, then we need to find out whether it is best to continue from $\langle C_1 = 1, X_1 = x_1\rangle$ or $\langle C_1 = 2, X_1 = x_1\rangle$ or $\langle C_1 = 3, X_1 = x_1\rangle$.

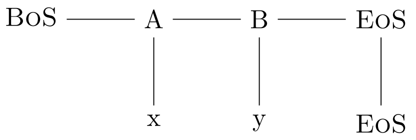
$$\alpha(2,1) = \max \begin{Bmatrix} \lambda_1^{(0)} \times \theta_{x_1}^{(1)} & \times & \lambda_1^{(1)} \times \theta_{x_2}^{(1)} \\ \lambda_2^{(0)} \times \theta_{x_1}^{(2)} & \times & \lambda_1^{(2)} \times \theta_{x_2}^{(1)} \\ \lambda_3^{(0)} \times \theta_{x_1}^{(3)} & \times & \lambda_1^{(3)} \times \theta_{x_2}^{(1)} \end{Bmatrix}$$

14. Consider the tagged sequences below where the **first sequence occurs n_1 times**, the **second sequence occurs n_2 times**, and the **third sequence occurs n_3 times**.

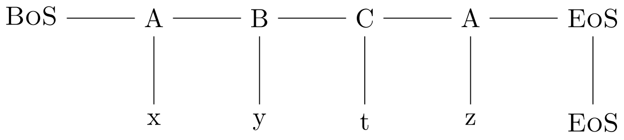
(1)



(2)



(3)



(a) (1 point) Estimate by maximum likelihood the transition distribution given that the previous category is 'A'.

Solution:

	C			
C_{prev}	A	B	C	EoS
A	0	$\frac{n_2+n_3}{n_1+n_2+2n_3}$	0	$\frac{n_1+n_3}{n_1+n_2+2n_3}$

(b) (1 point) Estimate by maximum likelihood the emission distribution given that the category is 'A'.

Solution:

	X			
C	t	x	y	z
A	0	$\frac{n_1+n_2+n_3}{n_1+n_2+2n_3}$	0	$\frac{n_3}{n_1+n_2+2n_3}$

(c) (1 point) What is the probability of the second sequence pair, given its length, as a function of maximum likelihood estimates?

Solution: BoS is only followed by 'A' in this dataset, thus the first transition has probability 1. Then we emit 'x' from 'A'. Then we transit to 'B' from 'A', then we emit 'y' from 'B' (which has probability 1 because 'B' is unambiguous in this dataset). Then we transit from 'B' to EoS. And, finally, we emit EoS from EoS with probability

1 because EoS is unambiguous by design.

$$\begin{aligned}
P_{X_1^n C_1^n|N}(\langle \text{A, B, EoS}\rangle, \langle \text{x, y, EoS}\rangle | n) &= P_{C|C_{\text{prev}}}(\text{A}|\text{BoS}) P_{X|C}(\text{x}|\text{A}) \\
&\times P_{C|C_{\text{prev}}}(\text{B}|\text{A}) P_{X|C}(\text{y}|\text{B}) \\
&\times P_{C|C_{\text{prev}}}(\text{EoS}|\text{B}) P_{X|C}(\text{EoS}|\text{EoS}) \\
&= 1 \times \frac{n_1 + n_2 + n_3}{n_1 + n_2 + 2n_3} \\
&\times \frac{n_2 + n_3}{n_1 + n_2 + 2n_3} \times 1 \\
&\times \frac{n_2}{n_2 + n_3} \times 1
\end{aligned}$$

Total for Question 14: 3

5 Probabilistic context-free grammars

Let $\mathfrak{G} = \langle \Sigma, \mathcal{V}, \mathrm{S}, \mathcal{R} \rangle$ be a context-free grammar (CFG) where

- Σ is the set of terminals
- $\mathcal{V}$ is the set of nonterminals
- $\mathrm{S} \in \mathcal{V}$ is the start symbol
- $\mathcal{R}$ is a set of context-free rules

and also assume that the most complex rule has a sequence of a symbols on its right-hand side (RHS).

15. (1 point) What is the general form of a context-free rule in $\mathcal{R}$? Make sure to formally specify the set to which left-hand side (LHS) and RHS belong.

> **Solution:** $\mathrm{X} \to \alpha$ where $\mathrm{X} \in \mathcal{V}$ is a nonterminal and $\alpha \in (\Sigma \cup \mathcal{V})^a$ is a possibly empty sequence of terminals and nonterminals

16. (1 point) If we know that $\mathfrak{G}$ is in Chomsky normal form (CNF), what can we say about rules in $\mathcal{R}$?

> **Solution:** Then every rule in $\mathcal{R}$ is either binary, unary, or $\mathrm{S} \to \epsilon$. Binary rules are of the form $\mathrm{A} \to \mathrm{B}\,\mathrm{C}$ where $\mathrm{A}, \mathrm{B}, \mathrm{C} \in \mathcal{V}$ are nonterminals. Unary rules are of the form $\mathrm{T} \to \mathrm{x}$ where $\mathrm{T} \in \mathcal{V}$ is a part-of-speech category (a nonterminal) and $\mathrm{x} \in \Sigma$ is a word (a terminal).

17. A probabilistic CFG (PCFG) extends a CFG with a probability distribution over derivations.

(a) (1/2 point) Define a random rule.

> **Solution:** A random rule $R = \langle v, \beta \rangle$ is a random pair where $v \in \mathcal{V}$ is a random LHS nonterminal and $\beta \in (\Sigma \cup \mathcal{V})^a$ is a random RHS sequence of terminals and nonterminals.

(b) (1/2 point) Define a random derivation.

> **Solution:** A random derivation D is a sequence $\langle R_1, \ldots, R_m \rangle$ of random rule applications, where R is a random variable that corresponds to rules in $\mathcal{R}$. A random derivation corresponds to a depth-first expansion of nonterminals starting from the start symbol until only terminal symbols remain.

(c) (1 point) Write down the probability distribution of a derivation r_1^m given its length as a function of the factor $P_{\mathrm{RHS}|\mathrm{LHS}}$.

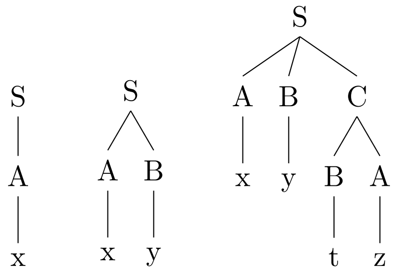
Solution: In a PCFG we make a 0-order Markov assumption, that is, rules are generated independently from the same rule distribution P_R. This means

$$P_{D|N}(r_1^m|m) = \prod_{i=1}^{m} P_R(r_i) = \prod_{i=1}^{m} P_{\mathrm{RHS|LHS}}(\beta_i|v_i)$$

where the ith rule r_i corresponds to $v_i \to \beta_i$.

Total for Question 17: 2

18. Consider the treebank below where the **first tree occurs n_1 times**, the **second tree occurs n_2 times**, and the **third tree occurs n_3 times**.



(a) (1 point) Use this treebank to derive the minimal set of context-free rules that could reconstruct it.

Solution:
S → A
S → A B
S → A B C
C → B A
A → x
A → z
B → y
B → t

(b) (1 point) Consider we extend this grammar to a PCFG, write down the maximum likelihood estimates for all pre-terminal rules.

Solution:

Rule	MLE
A → x	$\frac{n_1+n_2+n_3}{n_1+n_2+2n_3}$
A → z	$\frac{n_3}{n_1+n_2+2n_3}$
B → y	$\frac{n_2+n_3}{n_2+2n_3}$
B → t	$\frac{n_3}{n_2+2n_3}$

(c) (1 point) Write down a derivation (as an ordered sequence of rule applications) for the second tree.

Solution: A derivation is a depth-first traversal of the tree, therefore we have

$$\langle \mathrm{S} \rightarrow \mathrm{A\,B}, \mathrm{A} \rightarrow \mathrm{x}, \mathrm{B} \rightarrow \mathrm{y} \rangle$$

(d) (1 point) What is the probability of the second tree under a PCFG estimated via maximum likelihood using the given treebank.

Solution:

$$\begin{aligned}P_D(\langle \mathrm{S} \to \mathrm{A\,B}, \mathrm{A} \to \mathrm{x}, \mathrm{B} \to \mathrm{y}\rangle) &= P_{\mathrm{RHS|LHS}}(\mathrm{A\,B}|\mathrm{S}) \times P_{\mathrm{RHS|LHS}}(\mathrm{x}|\mathrm{A}) \times P_{\mathrm{RHS|LHS}}(\mathrm{y}|\mathrm{B})\\ &= \frac{n_2}{n_1+n_2+n_3} \times \frac{n_1+n_2+n_3}{n_1+n_2+2n_3} \times \frac{n_2+n_3}{n_2+2n_3}\end{aligned}$$

Total for Question 18: 4

6 Deductive systems

19. In the HMM model we often need to represent the space of all possible analyses of a sentence x_1^n, for example, we need that space of options in order to characterise the marginal probability $P_{X_1^n}(x_1^n)$ as well as in order to find the best tag sequence. Below, we have a deductive system that compactly represents the set of all possible analyses.

$$\begin{array}{rl}
\textsc{Input} & \text{tagset } \{1, \ldots, t\} \text{ and sentence } x_1^n \\
\textsc{Item} & [c, i] \quad \text{where } c \in \{1, \ldots, t\} \cup \{\textsc{BoS}, \textsc{EoS}\} \text{ and } i \in \{0, n+1\} \\
\textsc{Goal} & [\textsc{EoS}, n+1] \\
\textsc{Axioms} & [\textsc{BoS}, 0] \\
\textsc{Tag} & \dfrac{[c, i]}{[c', i+1]} \quad i < n \text{ and } c' \in \{1, \ldots, t\} \\
\textsc{Conclude} & \dfrac{[c, n]}{[\textsc{EoS}, n+1]}
\end{array}$$

In this program an item $[c, i]$ refers to word x_i being tagged with tag c. We augment the tag set $\{1, \ldots, t\}$ with two special symbols $\{\textsc{BoS}, \textsc{EoS}\}$ which help us track the beginning and the end of the tag sequence.

(a) (1 point) How many items can we prove for an input x_1^n (use big-O-notation)?

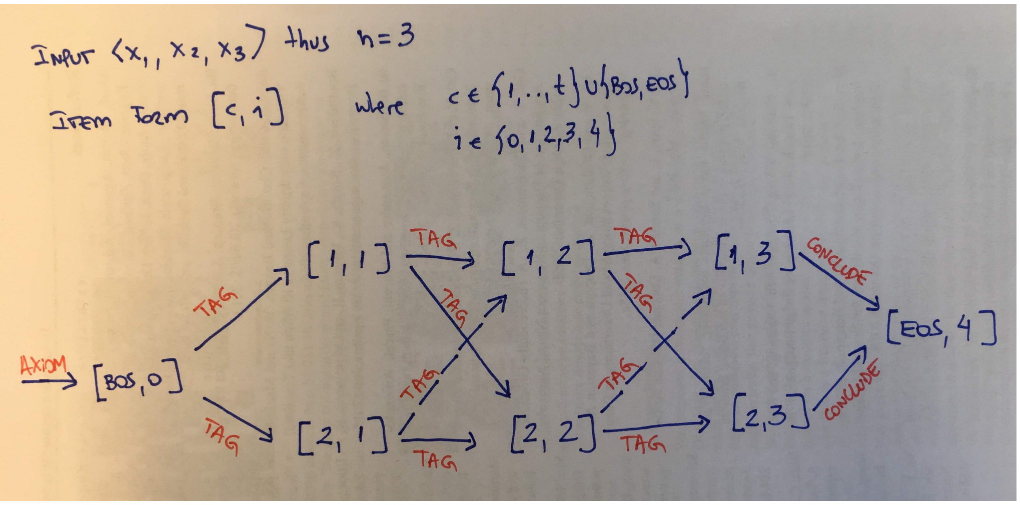
Solution: We have one item per tag-position pair, thus $O(t \times n)$.

(b) (1 point) How many inferences are valid for an input x_1^n (use big-O-notation)?

Solution: The most complex rule is TAG which leads to t inferences per item, as we have at $O(t \times n)$ items, we get at most $O(t^2 \times n)$ inferences.

(c) (2 points) For an input sentence $\langle x_1, x_2, x_3 \rangle$ and with $t = 2$, draw a graph where each state corresponds to an instantiation of a valid item and where each inference corresponds to an edge. Indicate which rule of the deductive system supports each edge in this graph.

Solution:



INPUT $\langle x_1, x_2, x_3 \rangle$ thus $n=3$

ITEM FORM $[c, i]$ where $c \in \{1, .., t\} \cup \{BOS, EOS\}$
$i \in \{0, 1, 2, 3, 4\}$

(d) (2 points) Each path of execution of the deductive system corresponds to one complete analysis, we can call it a *derivation* since it stands for a way to *derive* or *prove* the goal item. We can easily extend the system to assign a weight to each inference. Let us assume a parameterisation of our HMM generative story

$$C_i|C_{i-1} = c_{i-1} \sim \mathrm{Cat}(\lambda_1^{(c_{i-1})}, \ldots, \lambda_t^{(c_{i-1})})$$
$$X_i|C_i = c_i \sim \mathrm{Cat}(\theta_1^{(c_i)}, \ldots, \theta_v^{(c_i)})$$

in terms of transition and emission distributions (ignoring length). An Axiom is a trivial inference, thus we give it a dummy weight (not to interfere with the total)

$$\textsc{Axioms} \quad [\textsc{BoS}, 0] : \bar{1}$$

Conclude is an inference which serves the purpose of ending the tag sequence and the word sequence with special EoS tokens positioned at $i = n + 1$. This requires a transition to EoS and an emission of EoS, thus the weight is

$$\textsc{Conclude} \quad \frac{[c, n]}{[\textsc{EoS}, n+1] : \lambda_{\textsc{EoS}}^{(c)} \times \theta_{\textsc{EoS}}^{(\textsc{EoS})}}$$

What is the weight of the Tag rule?

Solution: The Tag rule transits from $C_i = c$ to $C_{i+1} = c'$ and emits $X_{i+1} = x_{i+1}$ which has probability $\lambda_{c'}^{(c)} \times \theta_{x_{i+1}}^{(c')}$, therefore

$$\textsc{Tag} \quad \frac{[c, i]}{[c', i+1] : \lambda_{c'}^{(c)} \times \theta_{x_{i+1}}^{(c')}} \quad i < n \text{ and } c' \in \{1, \ldots, t\}$$

Total for Question 19: 6

7 Misc

20. (2 points) Compare n-gram LMs, hidden Markov models, and probabilistic CFGs in terms of what they can or cannot capture, cost of representation, and algorithmic complexity of assessing important quantities such as marginals and argmax.

Solution: All these models can assign probability to sentences. n-gram LMs generate sentences one word at a time conditioning on a shortened history made of the last $n-1$ words. HMMs do not condition directly on the history of already generated words, instead a word is generated conditioned on its part-of-speech class, a PoS class on the other hand is generated based on the class of the preceding word. In PCFGs, the notion of linear order is replaced by a hierarchical organisation of abstract classes, and generation becomes a top-down story. n-gram LMs can only account for word order by memorising different sequences of conditioning contexts, thus it is very memory-inefficient requiring $O(v^n)$ parameters (where v is the size of the vocabulary). HMMs are much more memory-efficient, they only require $O(t^2 + t \times v)$ parameters because all of the memory is compressed in POS tags. PCFGs decompose in terms of context-free rule applications and the model factorises over pairs of LHS and RHS strings. For a CFG in CNF form, we have $O(|\mathcal{V}|^3 + t \times v)$ where $|\mathcal{V}|$ is the size of the nonterminal set, t is the number of pre-terminals (nonterminals that correspond to PoS tags) and v is the size of the vocabulary of terminals. We can thus see that PCFGs have more parameters than HMMs, that is the case because pre-terminal rules of a CFG correspond roughly to emissions in an HMM, and $|\mathcal{V}| > t$ (since the nonterminal set must at least include all parts-of-speech). On the other hand, without a huge increase in parameter space, a PCFG can relate events that happen very far apart (which n-gram and HMM LMs cannot). For example in a rule S $\rightarrow$ NP VP NP we get to relate two noun phrases that may end up far apart (if the verb phrase is a large phrase, for example). The only "downside" of HMMs and PCFGs is that we need to marginalise structure (tag sequence, or CFG trees) in order to get to the marginal probability of a sentence, but we have derived efficient (polynomial-time) algorithms for that, e.g. the Inside algorithm. These algorithms are based on the principle of dynamic programming whereby we decompose a complex problem recursively and use memoization to store previously computed solutions. HMMs and PCFGs can also be used to predict analyses for the sentences (e.g. tag sequences and CFG trees), using the Viterbi algorithm, which can be useful on their own right.

Assessment

Question	Points	Score
1	1	
2	1	
3	3	
4	1	
5	2	
6	1	
7	4	
8	3	
9	3	
10	2	
11	1	
12	2	
13	2	
14	3	
15	1	
16	1	
17	2	
18	4	
19	6	
20	2	
Total:	45	