Final for COMS W4705
Name:

30	15	15	30	15	15	30	20

Part #1 30 points

Consider a very simple bigram language model, where the vocabulary consists of the single word a, and the parameters of the model are

$$\begin{aligned} q(\texttt{a}|*) &= 1.0 \\ q(\texttt{a}|\texttt{a}) &= 0.4 \\ q(\texttt{STOP}|\texttt{a}) &= 0.6 \end{aligned}$$

Question 1 (5 points) What probabilities are assigned to the strings

a

a a

and

a a a

?

Question 2 (5 points) What probability is assigned to a string of n a's, where $n \geq 1$ (write the probability as a function of n).

Question 3 (5 points) Write down a probabilistic context-free grammar (PCFG) that defines the same distribution over strings as the language model above.

Question 4 (5 points) Write down a hidden Markov model (HMM) that defines the same distribution over strings as the language model above.

Question 5 (5 points) Now consider a bigram language model where the vocabulary consists of two words `a` and `b`, and the parameters of the model are

$$\begin{aligned} q(\texttt{a}|*) &= 0.5 \\ q(\texttt{b}|*) &= 0.5 \\ q(\texttt{a}|\texttt{a}) &= 0.2 \\ q(\texttt{b}|\texttt{a}) &= 0.2 \\ q(\texttt{a}|\texttt{b}) &= 0.2 \\ q(\texttt{b}|\texttt{b}) &= 0.2 \\ q(\texttt{STOP}|\texttt{a}) &= 0.6 \\ q(\texttt{STOP}|\texttt{b}) &= 0.6 \end{aligned}$$

Write down a PCFG that defines the same distribution over strings as this language model.

Question 6 (5 points) Write down an HMM that defines the same distribution over strings as the language model from the previous question.

Part #2 15 points

We define the following type of "lexicalized" grammar:

- N is a set of non-terminal symbols
- Σ is a set of terminal symbols
- R is a set of rules which take one of two forms:
  - $X(h) \rightarrow Y_1(h)\ Y_2(w)$ for $X \in N$, and $Y_1, Y_2 \in N$, and $h, w \in \Sigma$
  - $X(h) \rightarrow h$ for $X \in N$, and $h \in \Sigma$
- $S \in N$ is a distinguished start symbol

Note that this is similar to the "lexicalized Chomsky normal form" grammar we introduced in lecture, **except** that we do not allow rules of the following form:

$$X(h) \rightarrow Y_1(w)\ Y_2(h) \text{ for } X \in N, \text{ and } Y_1, Y_2 \in N, \text{ and } h, w \in \Sigma.$$

Question 7 (15 points)

Define a grammar in the above form that gives at least one valid parse tree for the sentence *the man and the man saw the man*. Draw a parse tree under your grammar for this sentence. Make sure to show the head words in your parse tree.

(This page left intentionally blank)

Part #3 15 points

Question 8 (15 points)

Nathan L. Pedant decides to build a trigram language model. He randomly selects 1000 sentences from the New York Times as training data, and randomly selects an additional (different) 1000 sentences from the New York Times as test data. He estimates the parameters of the trigram model as

$$p(w_3|w_1, w_2) = \frac{Count(w_1, w_2, w_3)}{Count(w_1, w_2)} \quad \text{if } Count(w_1, w_2) > 0$$

and

$$p(w_3|w_1, w_2) = \frac{1}{N} \quad \text{if } Count(w_1, w_2) = 0$$

where the $Count$ values are taken from the training data, N is the number of words in the vocabulary, and $p(w_3|w_1, w_2)$ is the probability of seeing w_3 given that the previous two words were w_1 and w_2. Note that this is an *unsmoothed* trigram model, where the parameters are simple maximum-likelihood estimates.

Nathan then measures the perplexity of his model on the test corpus.

Question: The perplexity on the test corpus will almost certainly be **infinity**. Specify precise conditions on the training/test set pair under which the perplexity on the test corpus for Nathan's model will be **finite**, and argue for why these conditions are unlikely to happen in practice.

(This page left intentionally blank)

Part #4 (30 points)

Question 9 (10 points) Consider the following problem concerning the IBM models for machine translation. We observe the following training examples:

$$\begin{aligned} e^{(1)} &= \text{the dog barks} & f^{(1)} &= \text{adog athe abarks} \\ e^{(2)} &= \text{the cat barks} & f^{(2)} &= \text{acat athe abarks} \\ e^{(3)} &= \text{a cat barks} & f^{(3)} &= \text{acat aa abarks} \end{aligned}$$

Write down the parameters of an IBM model 2 model, such that $p(f^{(i)}|e^{(i)}) = 1$ for $i \in \{1, 2, 3\}$ (for simplicity assume that English and French sentences are always of length 3).

Question 10 (20 points) Now consider an example where the training data is as follows:

$e^{(1)}$ = the dog barks	$f^{(1)}$ = abarks adog athe
$e^{(2)}$ = the dog barks	$f^{(2)}$ = athe adog abarks
$e^{(3)}$ = the cat barks	$f^{(3)}$ = abarks acat athe
$e^{(4)}$ = the cat barks	$f^{(4)}$ = athe acat abarks
$e^{(5)}$ = a cat barks	$f^{(5)}$ = abarks acat aa
$e^{(6)}$ = a cat barks	$f^{(6)}$ = aa acat abarks

Consider a translation model where for any French sentence $f_1 f_2 f_3$, English sentence $e_1 e_2 e_3$, and alignment variables $a_1 a_2 a_3$, we have

$$p(f_1 f_2 f_3, a_1 a_2 a_3 | e_1 e_2 e_3) = \prod_{j=1}^{3} d(a_j | a_{j-1}) \prod_{j=1}^{3} t(f_j | e_{a_j})$$

here $d(a_j|a_{j-1})$ is an alignment parameter, which is conditioned on the previous alignment variable. We define $a_0 = *$, where * is a special start symbol.

Define d and t parameters of a model of this form, such that $p(f^{(i)}|e^{(i)}) = 0.25$ for $i \in \{1, 2, \ldots, 6\}$ (for simplicity assume that English and French sentences are always of length 3).

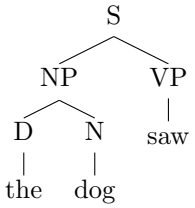
Part #5 (15 points)

In this question we develop a global linear model for parsing with a context-free grammar in Chomsky normal form. The input to the model is a sentence $x_1 \dots x_n$ where x_i is the i'th word in the sentence. We use $\mathcal{T}(x_1 \dots x_n)$ to denote the set of all parse trees for the sentence $x_1 \dots x_n$. For any parse tree $y \in \mathcal{T}(x_1 \dots x_n)$, for any rule $X \rightarrow Y\ Z$ in the grammar, for any indices i, k, j such that $1 \leq i \leq k < j \leq n$, we define

$$\delta(y, X \rightarrow Y\ Z, i, k, j) = 1$$

if the rule $X \rightarrow Y\ Z$ is seen in the parse tree y, with non-terminal X spanning words $i \dots j$ inclusive; non-terminal Y spanning words $i \dots k$ inclusive; and non-terminal Z spanning words $(k+1) \dots j$ inclusive.

For example, for the parse tree



we have $\delta(\text{S} \rightarrow \text{NP VP}, 1, 2, 3) = \delta(\text{NP} \rightarrow \text{D N}, 1, 1, 2) = 1$, with all other δ values being equal to 0.

We also assume that we have a feature vector $g(x_1 \dots x_n, X \rightarrow Y\ Z, i, k, j) \in \mathbb{R}^d$ for any sentence $x_1 \dots x_n$ together with a rule $X \rightarrow Y\ Z, i, k, j$; and a parameter vector $v \in \mathbb{R}^d$. The score for an entire parse tree under parameter values v is

$$f(y; v) = \sum_{X \rightarrow Y\ Z, i, k, j} \delta(y, X \rightarrow Y\ Z, i, k, j)\, (v \cdot g(x_1 \dots x_n, X \rightarrow Y\ Z, i, k, j))$$

Thus the score for an entire parse tree is a sum of scores for the rules it contains, where each rule receives the score

$$v \cdot g(x_1 \dots x_n, X \rightarrow Y\ Z, i, k, j)$$

Question 11 (15 points) Assume that the underlying context-free grammar is as follows:

S → NP VP	IN → of
NP → DT NN	IN → with
NP → NP PP	DT → the
VP → VP PP	NN → man
PP → IN NP	NN → telescope
VP → VB NP	VB → saw

Now assume that we have training data consisting of examples $x^{(i)}, y^{(i)}$ for $i = 1 \ldots n$ where each $x^{(i)}$ is a sentence, and each $y^{(i)}$ is a parse tree for the sentence that is valid under the above grammar. All parse trees in the training set satisfy the following constraints:

- Whenever the preposition *of* is seen in a sentence, it is the first word of a `PP` that modifies an `NP` (i.e., the `PP` is on the right hand side of a rule `NP → NP PP`).
- Whenever the preposition *with* is seen in a sentence, it is the first word of a `PP` that modifies a `VP` (i.e., the `PP` is on the right hand side of a rule `VP → VP PP`).

Give a definition of the feature vector $g(x_1 \ldots x_n, X \rightarrow Y\ Z, i, k, j)$ that will allow us to correctly model the training data given above.

Part #6 (15 points)

Consider an application of global linear models to dependency parsing. In this scenario each input $x = x_1 \dots x_n$ is a sentence. GEN$(x_1 \dots x_n)$ returns the set of all dependency parses for x. The feature vector $\mathrm{f}(x, y)$ for any sentence x paired with a dependency parse tree y is defined as

$$\mathrm{f}(x, y) = \sum_{(h,m) \in y} g(x, h, m)$$

where g is a function that maps a dependency (h, m) together with the sentence x to a local feature vector. Here h is the index of the head-word of the dependency, and m is the index of the modifier word.

Question 12 (15 points) Now assume that we have training data consisting of examples $x^{(i)}, y^{(i)}$ for $i = 1 \dots n$ where each $x^{(i)}$ is a sentence, and each $y^{(i)}$ is a dependency parse tree for the sentence. All dependency parse trees in the training set satisfy the following constraints:

- All dependencies have a head word that has an odd number of letters, and a modifier word that has an even number of letters.
- All dependencies have the head word to the left of the modifier word in the sentence.

Give a definition of the feature vector $g(x_1 \dots x_n, h, m)$ that will allow us to correctly model the training data given above.

Part #7 (30 points)

Clarissa decides to build a log-linear model for language modeling. She has a training sample (x_i, y_i) for $i = 1 \ldots n$, where each x_i is a word corresponding to the previous word in a document (e.g., $x_i =$ `said`) and y_i is the next word seen after this word (e.g., $y_i =$ `that`). As usual in log-linear models, she defines a function $\mathrm{f}(x, y)$ that maps any x, y pair to a vector in $\mathbb{R}^d$. Given parameter values $\theta \in \mathbb{R}^d$, the model defines

$$p(y|x;\theta) = \frac{e^{\theta \cdot \mathrm{f}(x,y)}}{\sum_{y' \in \mathcal{V}} e^{\theta \cdot \mathrm{f}(x,y')}}$$

where $\mathcal{V}$ is the *vocabulary*, i.e., the set of possible words; and $\theta \cdot \mathrm{f}(x, y)$ is the inner product between the vectors θ and $\mathrm{f}(x, y)$.

Given the training set, the training procedure returns parameters $\theta^* = \arg\max_\theta L(\theta)$, where

$$L(\theta) = \sum_i \log p(y_i|x_i;\theta) - \frac{\lambda}{2} \sum_k \theta_k^2$$

and $\lambda > 0$ is some constant.

Recall that for any parameter θ_j, we have

$$\frac{dL(\theta)}{d\theta_j} = \sum_{i=1}^{n} f_j(x_i, y_i) - \sum_{i=1}^{n} \sum_y p(y|x_i;\theta) f_j(x_i, y) - \lambda\theta_j$$

Clarissa makes the following choice of her first three features in the model:

$$f_1(x,y) = \begin{cases} 1 & \text{if } y = \texttt{model} \text{ and } x = \texttt{the} \\ 0 & \text{otherwise} \end{cases}$$

$$f_2(x,y) = \begin{cases} 1 & \text{if } y = \texttt{model} \text{ and } x = \texttt{the} \\ 0 & \text{otherwise} \end{cases}$$

$$f_3(x,y) = \begin{cases} 1 & \text{if } y = \texttt{model} \text{ and } x = \texttt{the} \\ 0 & \text{otherwise} \end{cases}$$

So $f_1(x, y)$, $f_2(x, y)$ and $f_3(x, y)$ are *identical features*.

(Question continues over the page.)

Question 13 (10 points) True or false? For any training set consisting of examples (x_i, y_i) for $i = 1 \ldots n$, where there is no training example i such that $x_i =$ `the` and $y_i =$ `model`, with f_1, f_2 and f_3 defined as above, the optimal parameters θ^* satisfy the property that $\theta_1^* = \theta_2^* = \theta_3^* = 0$. (Answer True or False below. Make sure to give a justification for your answer; you will only receive 5 points for a correct answer with an incorrect justification)

Question 14 (10 points) True or false? For any training set consisting of examples (x_i, y_i) for $i = 1 \ldots n$, where there is no training example i such that $x_i = \texttt{the}$, with f_1, f_2 and f_3 defined as above, the optimal parameters θ^* satisfy the property that $\theta_1^* = \theta_2^* = \theta_3^* = 0$. (Answer True or False below. Make sure to give a justification for your answer; you will only receive 5 points for a correct answer with an incorrect justification)

Question 15 (10 points) Now say we define the optimal parameters to be

$$\theta^* = \arg\max_\theta L(\theta)$$

where

$$L(\theta) = \sum_i \log p(y_i|x_i;\theta) - \frac{\lambda}{2}\sum_k \theta_k^4$$

True or false? For any training set, with f_1, f_2 and f_3 defined as above, the optimal parameters θ^* satisfy the property that $\theta_1^* = \theta_2^* = \theta_3^*$. (Answer True or False below. Make sure to give a justification for your answer; you will only receive 5 points for a correct answer with an incorrect justification)

Part #8 (20 points)

Say we are running the perceptron algorithm. We have reached example x_i and the set $\{f(x_i, y) : y \in \text{GEN}(x_i)\}$ consists of the following vectors:

(a) $\langle 1, 0, 0, 1 \rangle$

(b) $\langle 1, 1, 0, 0 \rangle$

(c) $\langle 0, 1, 1, 1 \rangle$

Assume also that $f(x_i, y_i) = \langle 1, 1, 0, 0 \rangle$.

Question 16 (10 points) Give a setting for the parameter vector v that ensures that the output of the global linear model on x_i is y_i.

Question 17 (10 points) Now assume that $v = \langle 0, 1, 0, 1 \rangle$ immediately before this example is considered by the algorithm. What will the value of v after the update on this example? (Make sure to give a full justification of your answer.)