LING/COMP 445, LING 645
Problem Set 5

Name: Jan Tiegges , **McGill ID**: 261180937
Collaborators:

Due before 4:35 PM on Wednesday, November 22, 2023

Please enter your name and McGill ID above.

This problem set consists only of questions involving mathematics or English or or a combination of the two (no coding questions this time). Please put your answers in an answer box like in the example below.

Once you have entered your answers, please compile your copy of this LaTeX into a PDF and submit

(i) the compiled PDF renamed to `ps5-lastname-firstname.pdf` and
(ii) the raw LaTeX file renamed to `ps5-lastname-firstname.tex`

to the Problem Set 5 folder under 'Assignments' on MyCourses.

Example Problem: This is an example question using some fake math like this $L = \sum_0^\infty \mathcal{G}\delta_x$.

Example Answer: Put your answer in the box provided, like this:

Example answer is $L = \sum_0^\infty \mathcal{G}\delta_x$.

Problem 1: In class we gave the following equation for the bigram probability of a sequence of words $W^{(1)}, \ldots, W^{(k)}$:

$$\Pr(W^{(1)}, \ldots, W^{(k)}) \stackrel{\text{def}}{=} \prod_{i}^{k} \Pr(W^{(i)} | W^{(i-1)} = w^{(i-1)}) \tag{1}$$

Using this formula, give an expression for the bigram probability of the sentence *abab*, where each character is treated as a word. Try to simplify the formula as much as possible.

Important note: Throughout this problem set, the vocabulary will be $V \stackrel{\text{def}}{=} \{a, b\}$. We will assume the length of the sentence is fixed at some k, and *we will not use the stop symbol.* That is, in a sentence of length k, for $1 \leq i \leq k$, the possible values for the random variable $W^{(i)}$ are just a and b, and we will refer to the beginning of the string as $W^{(0)} = \rtimes$ always. So, $\Pr(W^{(1)} = a \mid W^{(0)} = \rtimes)$ is the probability that the string starts with a.

Answer 1: Please put your answer in the box below.

$$\Pr(W^{(1)} = a \mid W^{(0)} = \rtimes) \times \Pr(W^{(2)} = b \mid W^{(1)} = a) \times \Pr(W^{(3)} = a \mid W^{(2)} = b) \times \Pr(W^{(4)} = b \mid W^{(3)} = a)$$

Problem 2: There are two possible symbols/words in our language, a and b. There are three conditional distributions in the bigram model for this language, $\Pr(W^{(i)} | W^{(i-1)} = a)$, $\Pr(W^{(i)} | W^{(i-1)} = b)$, and $\Pr(W^{(i)} | W^{(i-1)} = \rtimes)$. These conditional distributions are associated with the parameter vectors $\vec{\theta}_a$, $\vec{\theta}_b$, and $\vec{\theta}_{\rtimes}$, respectively (these parameter vectors were implicit in the previous problem). For the current problem, we will assume that these parameters are fixed. Use a second subscript notation to denote components of these vectors, so $\theta_{ab} = \Pr(W^{(i)} = b \mid W^{(i-1)} = a)$.

Suppose that we are given a sentence $W^{(1)}, \ldots, W^{(k)}$. We will use the notation $n_{x \to y}$ to denote the number of times that the symbol y occurs immediately following the symbol x in the sentence. For example, $n_{a \to a}$ counts the number of times that symbol a occurs immediately following the symbol a. Using Equation 1, give an expression for the probability of a length k sentence in our language:

$$\Pr(W^{(1)}, \ldots, W^{(k)} | \vec{\theta}_a, \vec{\theta}_b, \vec{\theta}_{\rtimes})$$

The expression should make use of the $n_{x \to y}$ notation defined above.

Hint: the expression should be analogous to the formula that we found for the likelihood of a corpus under a bag of words model.

Answer 2: Please put your answer in the box below.

$$\Pr(W^{(1)}, \ldots, W^{(k)} \mid \vec{\theta}_a, \vec{\theta}_b, \vec{\theta}_{\rtimes}) = \theta_{\rtimes W^{(1)}} \times \prod_{x \in \{a,b,\rtimes\}} \prod_{y \in \{a,b\}} \theta_{xy}^{n_{x \to y}}$$

Problem 3: Assume the parameter vectors in our bigram model have the following values:

$$\vec{\theta}_a = (0.7, 0.3)$$
$$\vec{\theta}_b = (0.2, 0.8)$$
$$\vec{\theta}_{\rtimes} = (0.5, 0.5)$$

The first vector indicates that if the current symbol a, there is probability 0.7 of transitioning to the symbol a, and probability 0.3 of transitioning to the symbol b. Using your answer to the previous problem and these parameter values, calculate the probability of the string *aabab*.

Answer 3: Please put your answer in the box below.

$$\Pr(aabab) = \theta_{\rtimes a}^{n_{\rtimes \to a}} \times \theta_{aa}^{n_{a \to a}} \times \theta_{ab}^{n_{a \to b}} \times \theta_{ba}^{n_{b \to a}} = 0.5^1 \times 0.7^1 \times 0.3^2 \times 0.2^1 = 0.0063$$

Problem 4: In Problem 2, you found an expression for the bigram probability of a sentence in our language, which contains the symbols a and b. In that problem, we assumed that there were fixed parameter vectors $\vec{\theta}$ associated with each conditional distribution. In this problem, we will consider the setting in which we have uncertainty about the value of these parameters.

As we did in class, we will use the Dirichlet distribution to define a prior over parameters. Assume each parameter vector is drawn independently given $\vec{\alpha}$:

$$\vec{\theta}_{\mathbf{c}} \mid \vec{\alpha} \sim \text{Dirichlet}(\vec{\alpha}) \tag{2}$$
$$w^{(i)} \mid w^{(i-1)} \sim \text{categorical}(\vec{\theta}_{w^{(i-1)}}) \tag{3}$$
$$w^{(1)} \sim \text{categorical}(\vec{\theta}_{\rtimes}) \tag{4}$$

Suppose that we have a fixed-length sentence $S = W^{(1)}, \ldots, W^{(k)}$. Give an expression for the joint probability $\Pr(S, \vec{\theta}_a, \vec{\theta}_b, \vec{\theta}_{\rtimes} | \vec{\alpha})$ using the definitions of Dirichlet distributions and likelihoods we defined in class.

Hint: The joint probability $\Pr(S, \vec{\theta}_a, \vec{\theta}_b, \vec{\theta}_{\rtimes} \mid \vec{\alpha}) = \Pr(S \mid \vec{\theta}_a, \vec{\theta}_b, \vec{\theta}_{\rtimes}) \Pr(\vec{\theta}_a, \vec{\theta}_b, \vec{\theta}_{\rtimes} \mid \vec{\alpha})$, because S is conditionally independent of $\vec{\alpha}$ given the parameters. You solved $\Pr(S \mid \vec{\theta}_a, \vec{\theta}_b, \vec{\theta}_{\rtimes})$ in Problem 2.

Answer 4: Please put your answer in the box below.

$$\Pr(S, \vec{\theta}_a, \vec{\theta}_b, \vec{\theta}_{\rtimes} | \vec{\alpha}) = \Pr(S | \vec{\theta}_a, \vec{\theta}_b, \vec{\theta}_{\rtimes}) \times \Pr(\vec{\theta}_a | \vec{\alpha}) \times \Pr(\vec{\theta}_b | \vec{\alpha}) \times \Pr(\vec{\theta}_{\rtimes} | \vec{\alpha})$$

Expanding this (using Solution 2), we get:

$$\left(\prod_{x \in \{a,b,\rtimes\}} \prod_{y \in \{a,b\}} \theta_{xy}^{n_{x \to y}} \right) \times \left(\frac{1}{B(\vec{\alpha})} \prod_{i=1}^{D} \theta_{ai}^{\alpha_i - 1} \right) \times \left(\frac{1}{B(\vec{\alpha})} \prod_{i=1}^{D} \theta_{bi}^{\alpha_i - 1} \right) \times \left(\frac{1}{B(\vec{\alpha})} \prod_{i=1}^{D} \theta_{\rtimes i}^{\alpha_i - 1} \right)$$

Here, $B(\vec{\alpha})$ is the normalization constant for the Dirichlet distribution.

Problem 5: In the previous problem, you found a formula for the joint probability of a sentence and a set of bigram model parameters. Using this, give a formula for the marginal probability of the sentence $\Pr(S|\vec{\alpha})$.

Hint: The formula should be analogous to the formula derived in class for marginal probability of a corpus under a bag of words model. Whereas before there was only a single parameter vector $\vec{\theta}$, now there are three parameter vectors that need to be marginalized away. Otherwise the calculation will be similar.

Answer 5: Please put your answer in the box below.

$$\Pr(S|\vec{\alpha}) = \int\int\int \Pr(S, \vec{\theta}_a, \vec{\theta}_b, \vec{\theta}_{\rtimes}|\vec{\alpha})\, d\vec{\theta}_a\, d\vec{\theta}_b\, d\vec{\theta}_{\rtimes}$$

Problem 6: Let us assume that the parameters of the Dirichlet distribution are $\vec{\alpha} = (1, 1)$. Using your solution to the previous problem, write an expression for $\Pr(S = aabab \mid \vec{\alpha} = (1, 1))$, the marginal probability of the string *aabab*. The expression should should contain the gamma function $\Gamma(\cdot)$. Using the properties of the gamma function discussed in class (i.e., it's relationship to the factorial) or an online calculator, compute a numerical value for this expression.

Answer 6: Please put your answer in the box below.

The joint probability, including the Dirichlet distributions, is:

$$\Pr(S, \vec{\theta}_a, \vec{\theta}_b, \vec{\theta}_{\rtimes}|\vec{\alpha} = (1,1)) = \left(\theta_{\rtimes a} \times \theta_{aa} \times \theta_{ab}^2 \times \theta_{ba}\right) \times \frac{\Gamma(2)}{\Gamma(1)\Gamma(1)}^3 \times \prod_{\mathbf{c} \in \{a,b,\rtimes\}} \theta_{\mathbf{c}a}^0 \theta_{\mathbf{c}b}^0$$

The Gamma function $\Gamma(n)$ is equivalent to $(n-1)!$ for positive integers. Thus, $\Gamma(2) = 1!$ and $\Gamma(1) = 0!$. So the fraction $\frac{\Gamma(2)}{\Gamma(1)\Gamma(1)}$ becomes 1.

Integrating over the parameter vectors:

$$\Pr(S = aabab|\vec{\alpha} = (1,1)) = \int_0^1 \int_0^1 \int_0^1 \left(\theta_{\rtimes a} \times \theta_{aa} \times \theta_{ab}^2 \times \theta_{ba}\right)\, d\vec{\theta}_a\, d\vec{\theta}_b\, d\vec{\theta}_{\rtimes}$$

This integral simplifies the product of the expected values of each term under the uniform distribution from 0 to 1, since each θ_{xy} is independent and uniformly distributed.

Therefore, the numerical value is:

$$\Pr(S = aabab|\vec{\alpha} = (1,1)) = \frac{1}{2} \times \frac{1}{2} \times \left(\frac{1}{3}\right)^2 \times \frac{1}{2} = \frac{1}{48} \approx 0.02083$$

Problem 7: Suppose that we have observed a sentence $S = W^{(1)}, \ldots, W^{(k)}$. Find an expression for the posterior distribution over the model parameters, $\Pr(\vec{\theta}_a, \vec{\theta}_b, \vec{\theta}_{\rtimes} \mid S, \vec{\alpha})$.

Hint: Use the joint probability that you computed in Problem 4 and Bayes' rule. The solution should be analogous to the posterior probability for the bag of words model.

Answer 7: Please put your answer in the box below.

$$\Pr(\vec{\theta} \mid S, \vec{\alpha}) = \frac{\prod_{x \in \{a,b,\rtimes\}} \prod_{y \in \{a,b\}} \theta_{xy}^{n_{x \to y}} \times \prod_{c \in \{a,b,\rtimes\}} \text{Dirichlet}(\vec{\theta}_c; \vec{\alpha})}{\Pr(S \mid \vec{\alpha})}$$

Problem 8: Consider the language $L = \{a^*ba^*\}$, that is, the language consisting of some number of a's, followed by a single b, followed by some number of a's. Show that this language is not strictly 2-local.

Hint: use n-Local Suffix Substitution Closure (n-SSC).

Answer 8: Please put your answer in the box below.

According to 2-SCC, we should be able to pick any two strings s_1 and s_2 in L, and it must be the case that for all length 1 pivots x that are shared between s_1 and s_2, if $s_1 = u_1xv_1$ and $s_2 = u_2xv_2$ then u_1xv_2 is also a member of L. Let $s_1 = \rtimes abaa\ltimes$ with $u_1 = \rtimes ab$, $x = a$, and $v_1 = a\ltimes$. Let $s_2 = \rtimes aaba\ltimes$ with $u_2 = \rtimes a$, $x = a$, and $v_2 = ba\ltimes$.

Then, $u_1xv_2 = \rtimes abab\ltimes$ is not a member of L. Therefore, L is not strictly 2-local.

Problem 9: Consider the language $L = \{a^nb^mc^nd^m \mid n, m \in \mathbb{N}\}$, that is, the language consisting of n a's followed by m b's followed by n c's followed by m d's where n and m are natural numbers. Show that this language is not strictly 2-local.

Hint: use the same property as in the problem above.

Answer 9: Please put your answer in the box below.

Let $s_1 = \rtimes aabccd\ltimes$ with $u_1 = \rtimes aabc$, $x = c$, and $v_1 = d\ltimes$. Let $s_2 = \rtimes aaabcccd\ltimes$ with $u_2 = \rtimes aaabc$, $x = c$, and $v_2 = cd\ltimes$.

Then, constructing u_1xv_2 gives us $\rtimes aabcccd\ltimes$. However, this string is not a member of L, as the number of a's does not match the number of c's. Therefore, L is not strictly 2-local.

Problem 10: Show that the language $L = \{a^nb^mc^nd^m \mid n, m \in \mathbb{N}\}$ is not strictly k-local, for any value of k.

Answer 10: Please put your answer in the box below.

In L, the number of a's must match the number of c's, and the number of b's must match the number of d's. This relationship cannot be captured by any fixed k-length substrings, as k is finite and the language's structure depends on potentially infinite relationships between a's and c's, and b's and d's.

Therefore, no matter how large k is, there will always be strings that are in the language but not recognized as such, or vice versa, based on the k-length substrings they contain. Hence, the language L is not strictly k-local for any value of k.

Problem 11: In class we proved that $L \in \mathrm{SL}_k \implies k\text{−SSC}(L)$. In other words, if a language is k-strictly local, then it satisfies k-Local Suffix Substitution Closure.

Use this theorem to prove that k-strictly local languages are closed under intersection. More precisely, prove that if $L_1 \in \mathrm{SL}_k$ and $L_2 \in \mathrm{SL}_k$, then $L_1 \cap L_2 \in \mathrm{SL}_k$.

Answer 11: Please put your answer in the box below.

Consider a string s in $L_1 \cap L_2$. Since s is in both L_1 and L_2, any k-length substring of s is also a k-length substring of both L_1 and L_2. Furthermore, for any k-length substring in $L_1 \cap L_2$, the k-SSC property for both L_1 and L_2 ensures that any string formed by keeping the k-length prefix constant and varying the suffix according to the rules of L_1 and L_2 will still belong to both languages.

Thus, $L_1 \cap L_2$ is defined by the set of k-length substrings that are common to both L_1 and L_2. Since these substrings satisfy the k-SSC in both languages individually, they will also satisfy it in the intersection. Hence, the intersection $L_1 \cap L_2$ adheres to the definition of a k-strictly local language and is therefore k-strictly local.

LONG FORM READING QUESTION:

(This section is optional for students in LING/COMP 445, but must be completed if taking LING 645.)

You must answer this question on your own.

Delétang et al. (2023) demonstrate the generative capacities of different neural network language models (RNNs, LSTMs, Transformers, and memory enhansed RNNs) by trying to situate them in practice along the Chomsky hierarchy. Large language models like GPT-3 and Bard are transformer-based architectures. On the surface, such models generate natural sounding language (Have you tried ChatGPT?). Given what you have learnt about these types of architectures from Delétang et al.'s experiments, do you think we should expect these models to have learnt grammatical systems for natural language that are similar to the generative systems we think people may have (think generative grammars)? Or are these models likely doing some different entirely? Motivate your answer with at least 2 observations supporting your position.

Answer: Please put your answer in the box below.

The experiments of Delétang et al. show that transformer-based architectures such as GPT-3 and Bard are limited in their ability to learn complex and highly structured generalization patterns. While these models can learn some counter languages, they cannot recognize certain regular languages, such as periodic finite languages.

The authors highlight a limitation of transformers resulting from the lack of ability to use an external memory structure. These models were only able to generalize to lengths close to those seen during training. They show that the ability to use an external memory structure enables a model to climb the Chomsky hierarchy, and that the use of an external memory structure significantly improves a model's ability to learn complex and highly structured generalization patterns.

The observations therefore suggest that transformer-based architectures such as GPT-3 and Bard have not learned grammatical systems for natural language in the way we assume humans have. Humans are able to learn complex and highly structured generalization patterns, and can recognize regular languages such as periodic finite languages. The fact that transformer-based architectures are unable to learn these patterns suggests that they are not learning in the same way as humans. Instead, they seem to rely on statistical patterns in the training data to generate plausible sentences without necessarily understanding the underlying rules for language use.