

LING/COMP 445, LING 645

Problem Set 5

Name: Mathieu-Joseph Magri , **McGill ID:** 260928498

Due before 8:35 AM on Thursday, March 30, 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 L^AT_EX into a PDF and submit

- (i) the compiled PDF renamed to **ps5-lastname-firstname.pdf** and
- (ii) the raw L^AT_EX 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 $W^{(1)}, \dots, W^{(k)}$:

$$\Pr(W^{(1)}, \dots, W^{(k)}) \stackrel{\text{def}}{=} \prod_i^k \Pr(W^{(i)} | W^{(i-1)} = w^{(i-1)}) \quad (1)$$

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 $W^{(0)} = \times$ always. So, $\Pr(W^{(1)} = a | W^{(0)} = \times)$ is the probability that the string starts with a .

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

We want to give the probability for the sentence *abab*, which, in explicit form, is the probability for $\times abab \times$. Therefore, the bigram probability for $\times abab \times$ can be written as follows:

$$\Pr(W^{(1)} = a | W^{(0)} = \times) * \Pr(W^{(2)} = b | W^{(1)} = a) * \Pr(W^{(3)} = a | W^{(2)} = b) * \Pr(W^{(4)} = b | W^{(3)} = a) * \Pr(W^{(5)} = \times | W^{(4)} = b)$$

The above equation can be simplified to:

$$\Pr(W^{(1)} = a | W^{(0)} = \times) * (\Pr(b | a))^2 * \Pr(W^{(3)} = a | W^{(2)} = b) * \Pr(W^{(5)} = \times | W^{(4)} = b)$$

Note: the usage of the $*$ symbol refers to a regular multiplication operator. It will be used to symbolise the multiplication operator for the entire assignment.

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 $\Pr(W^{(i)} | W^{(i-1)} = \times)$. These conditional distributions are associated with the parameter vectors $\vec{\theta}_a$, $\vec{\theta}_b$, and $\vec{\theta}_\times$, 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 $\theta_{ab} = \Pr(W^{(i)} = b | W^{(i-1)} = a)$.

Suppose that we are given a sentence $W^{(1)}, \dots, W^{(k)}$. We will use the notation $n_{x \rightarrow y}$ to denote the number of times that the symbol y occurs immediately following the symbol x in the sentence. For example, $n_{a \rightarrow a}$ 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:

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

The expression should make use of the $n_{x \rightarrow y}$ 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.

Using the hint, the formula that we found for the probability of a sentence under the bag of words model can be defined as:

$$\Pr(w^{(1)}, \dots, w^{(k)} \mid \vec{\theta}) = \prod_{w \in V} \theta_w^{n_w}$$

while the formula that we found for the likelihood of a corpus under the bag of words model can be defined as:

$$\Pr(C \mid \vec{\theta}) = \prod_{w \in V} \theta_w^{n_w}.$$

Using the previous formulas, we can write the expression for the probability of a length k sentence in our language as such:

$$\Pr(W^{(1)}, \dots, W^{(k)} \mid \vec{\theta}_a, \vec{\theta}_b, \vec{\theta}_{\times}) = \prod_{w \in V} \theta_{\times \rightarrow w}^{n_{\times \rightarrow w}} * \prod_{w \in V} \theta_{a \rightarrow w}^{n_{a \rightarrow w}} * \prod_{w \in V} \theta_{b \rightarrow w}^{n_{b \rightarrow w}} * \prod_{w \in V} \theta_{\times \rightarrow w}^{n_{\times \rightarrow w}}$$

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

$$\vec{\theta}_a = (0.7, 0.3)$$

$$\vec{\theta}_b = (0.2, 0.8)$$

$$\vec{\theta}_{\times} = (0.5, 0.5)$$

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.

We need to calculate the probability for $aabab$, or, in more explicit terms, $\times aabab \times$.

This should give us the following:

$$\theta_{\times \rightarrow a} * \theta_{a \rightarrow a} * \theta_{a \rightarrow b} * \theta_{b \rightarrow a} * \theta_{a \rightarrow b} * \theta_{b \rightarrow \times} = 0.5 * 0.7 * 0.3 * 0.2 * 0.3 * 0.5 = 0.00315 = 0.315\%$$

Note: We don't have the probability of a $\theta_{b \rightarrow \times}$ transition, therefore, I gave it the value of 0.5, the same as $\theta_{\times \rightarrow a}$ and $\theta_{\times \rightarrow b}$. If we don't consider the last transition, we end up with $0.5 * 0.7 * 0.3 * 0.2 * 0.3 = 0.0063 = 0.63\%$

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 $\vec{\alpha}$:

$$\vec{\theta}_c \mid \vec{\alpha} \sim \text{Dirichlet}(\vec{\alpha}) \tag{2}$$

$$w^{(i)} \mid w^{(i-1)} \sim \text{categorical}(\vec{\theta}_{w^{(i-1)}}) \tag{3}$$

$$w^{(1)} \sim \text{categorical}(\vec{\theta}_{\times}) \tag{4}$$

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

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

In class, we saw the Dirichlet distribution used to define the following:

$$p(\theta_1, \dots, \theta_K; \alpha_1, \dots, \alpha_K) = \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)} \prod_{i=1}^K \theta_i^{\alpha_i-1}$$

Furthermore, we know that $\Pr(C, \vec{\theta} | \vec{\alpha}) = \Pr(C | \vec{\theta}) \Pr(\vec{\theta} | \vec{\alpha})$ due to conditional independence.

In our situation, we need to find $\Pr(S, \vec{\theta}_a, \vec{\theta}_b, \vec{\theta}_\times | \vec{\alpha})$ using the Dirichlet distribution. We must do the following:

$$\begin{aligned} \Pr(S, \vec{\theta}_a, \vec{\theta}_b, \vec{\theta}_\times | \vec{\alpha}) &= \Pr(S | \vec{\theta}_a, \vec{\theta}_b, \vec{\theta}_\times) * \Pr(\vec{\theta}_a, \vec{\theta}_b, \vec{\theta}_\times | \vec{\alpha}) = \\ \Pr(S | \vec{\theta}_a) * \Pr(S | \vec{\theta}_b) * \Pr(S | \vec{\theta}_\times) * \Pr(\vec{\theta}_a | \vec{\alpha}) * \Pr(\vec{\theta}_b | \vec{\alpha}) * \Pr(\vec{\theta}_\times | \vec{\alpha}) &= \end{aligned}$$

$$\frac{\Gamma(\sum_{w \in V} \alpha_w)}{\prod_{w \in V} \Gamma(\alpha_w)} \prod_{w \in V} \theta_{a \rightarrow w}^{n_{a \rightarrow w} + \alpha_{a \rightarrow w} - 1} * \frac{\Gamma(\sum_{w \in V} \alpha_w)}{\prod_{w \in V} \Gamma(\alpha_w)} \prod_{w \in V} \theta_{b \rightarrow w}^{n_{b \rightarrow w} + \alpha_{b \rightarrow w} - 1} * \frac{\Gamma(\sum_{w \in V} \alpha_w)}{\prod_{w \in V} \Gamma(\alpha_w)} \prod_{w \in V} \theta_{\times \rightarrow w}^{n_{\times \rightarrow w} + \alpha_{\times \rightarrow w} - 1}$$

The above simplifies to:

$$\left(\frac{\Gamma(\sum_{w \in V} \alpha_w)}{\prod_{w \in V} \Gamma(\alpha_w)} \right)^3 \prod_{w \in V} \theta_{a \rightarrow w}^{n_{a \rightarrow w} + \alpha_{a \rightarrow w} - 1} * \prod_{w \in V} \theta_{b \rightarrow w}^{n_{b \rightarrow w} + \alpha_{b \rightarrow w} - 1} * \prod_{w \in V} \theta_{\times \rightarrow w}^{n_{\times \rightarrow w} + \alpha_{\times \rightarrow w} - 1}$$

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.

If we follow the procedure present in the course notes in Chapter 16, we should get the following:

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

$$\int_{\vec{\theta}_a} \Pr(S | \vec{\theta}_a) \Pr(\vec{\theta}_a | \vec{\alpha}) d\vec{\theta}_a * \int_{\vec{\theta}_b} \Pr(S | \vec{\theta}_b) \Pr(\vec{\theta}_b | \vec{\alpha}) d\vec{\theta}_b * \int_{\vec{\theta}_\times} \Pr(S | \vec{\theta}_\times) \Pr(\vec{\theta}_\times | \vec{\alpha}) d\vec{\theta}_\times =$$

$$\left(\frac{\Gamma(\sum_{w \in V} \alpha_w)}{\prod_{w \in V} \Gamma(\alpha_w)} \right)^3 \int_{\vec{\theta}_a} \prod_{w \in V} \theta_{a \rightarrow w}^{n_{a \rightarrow w} + \alpha_{a \rightarrow w} - 1} d\vec{\theta}_a * \int_{\vec{\theta}_b} \prod_{w \in V} \theta_{b \rightarrow w}^{n_{b \rightarrow w} + \alpha_{b \rightarrow w} - 1} d\vec{\theta}_b * \int_{\vec{\theta}_\times} \prod_{w \in V} \theta_{\times \rightarrow w}^{n_{\times \rightarrow w} + \alpha_{\times \rightarrow w} - 1} d\vec{\theta}_\times =$$

$$\left(\frac{\Gamma(\sum_{w \in V} \alpha_w)}{\prod_{w \in V} \Gamma(\alpha_w)}\right)^3 * \frac{\prod_{w \in V} \Gamma(n_{a \rightarrow w} + \alpha_{a \rightarrow w})}{\Gamma(\sum_{w \in V} (n_{a \rightarrow w} + \alpha_{a \rightarrow w}))} * \frac{\prod_{w \in V} \Gamma(n_{b \rightarrow w} + \alpha_{b \rightarrow w})}{\Gamma(\sum_{w \in V} (n_{b \rightarrow w} + \alpha_{b \rightarrow w}))} * \frac{\prod_{w \in V} \Gamma(n_{\text{!} \rightarrow w} + \alpha_{\text{!} \rightarrow w})}{\Gamma(\sum_{w \in V} (n_{\text{!} \rightarrow w} + \alpha_{\text{!} \rightarrow w}))}$$

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 contain the [gamma function](#) $\Gamma(\cdot)$. Using the properties of the gamma function discussed in class (i.e., its 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.

Replacing all values with from the answer of Question 5 with $\Pr(S = aabab \mid \vec{\alpha} = (1, 1))$, we get the following:

$$\begin{aligned} & \left(\frac{\Gamma(1+1)}{\Gamma(1)\Gamma(1)}\right)^3 * \frac{\Gamma(1+1)\Gamma(2+1)}{\Gamma(1+1+2+1)} * \frac{\Gamma(1+1)\Gamma(0+1)}{\Gamma(1+1+0+1)} * \frac{\Gamma(1+1)\Gamma(0+1)}{\Gamma(1+1+0+1)} = \\ & \left(\frac{\Gamma(2)}{1}\right)^3 * \frac{\Gamma(2)\Gamma(3)}{\Gamma(5)} * \frac{\Gamma(2)\Gamma(1)}{\Gamma(3)} * \frac{\Gamma(2)\Gamma(1)}{\Gamma(3)} = \\ & 1^3 * \frac{2}{24} * \frac{1}{2} * \frac{1}{2} = \frac{1}{48} \end{aligned}$$

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

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.

Bayes' rule is the following:

$$\Pr(\Theta = \vec{\theta} \mid \mathbf{C} = C, \vec{\alpha}) = \frac{\Pr(\mathbf{C} = C \mid \Theta = \vec{\theta}, \vec{\alpha}) \Pr(\Theta = \vec{\theta} \mid \vec{\alpha})}{\int_{\Theta} \Pr(\mathbf{C} = C \mid \Theta = \vec{\theta}, \vec{\alpha}) \Pr(\Theta = \vec{\theta} \mid \vec{\alpha})}$$

The numerator is the joint probability we calculated in Problem 4, while the denominator of the fraction is the marginal we calculated in problem 5. Therefore, replacing both, we get the following:

$$\frac{\left(\frac{\Gamma(\sum_{w \in V} \alpha_w)}{\prod_{w \in V} \Gamma(\alpha_w)}\right)^3 \prod_{w \in V} \theta_{a \rightarrow w}^{n_{a \rightarrow w} + \alpha_{a \rightarrow w} - 1} * \prod_{w \in V} \theta_{b \rightarrow w}^{n_{b \rightarrow w} + \alpha_{b \rightarrow w} - 1} * \prod_{w \in V} \theta_{\text{!} \rightarrow w}^{n_{\text{!} \rightarrow w} + \alpha_{\text{!} \rightarrow w} - 1}}{\left(\frac{\Gamma(\sum_{w \in V} \alpha_w)}{\prod_{w \in V} \Gamma(\alpha_w)}\right)^3 * \frac{\prod_{w \in V} \Gamma(n_{a \rightarrow w} + \alpha_{a \rightarrow w})}{\Gamma(\sum_{w \in V} (n_{a \rightarrow w} + \alpha_{a \rightarrow w}))} * \frac{\prod_{w \in V} \Gamma(n_{b \rightarrow w} + \alpha_{b \rightarrow w})}{\Gamma(\sum_{w \in V} (n_{b \rightarrow w} + \alpha_{b \rightarrow w}))} * \frac{\prod_{w \in V} \Gamma(n_{\text{!} \rightarrow w} + \alpha_{\text{!} \rightarrow w})}{\Gamma(\sum_{w \in V} (n_{\text{!} \rightarrow w} + \alpha_{\text{!} \rightarrow w}))}}$$

We can simplify the above to get:

$$\frac{\prod_{w \in V} \theta_{a \rightarrow w}^{n_{a \rightarrow w} + \alpha_{a \rightarrow w} - 1} * \prod_{w \in V} \theta_{b \rightarrow w}^{n_{b \rightarrow w} + \alpha_{b \rightarrow w} - 1} * \prod_{w \in V} \theta_{\times \rightarrow w}^{n_{\times \rightarrow w} + \alpha_{\times \rightarrow w} - 1}}{\frac{\prod_{w \in V} \Gamma(n_{a \rightarrow w} + \alpha_{a \rightarrow w})}{\Gamma(\sum_{w \in V} (n_{a \rightarrow w} + \alpha_{a \rightarrow w}))}} * \frac{\prod_{w \in V} \Gamma(n_{b \rightarrow w} + \alpha_{b \rightarrow w})}{\Gamma(\sum_{w \in V} (n_{b \rightarrow w} + \alpha_{b \rightarrow w}))}} * \frac{\prod_{w \in V} \Gamma(n_{\times \rightarrow w} + \alpha_{\times \rightarrow w})}{\Gamma(\sum_{w \in V} (n_{\times \rightarrow w} + \alpha_{\times \rightarrow w}))}}$$

Problem 8: Using your formula for the posterior distribution and setting $\vec{\alpha} = (1, 1)$, given an expression for the posterior distribution over parameters given the sentence *aabab* (don't evaluate it at particular parameter values). There should be an easy way to interpret the posterior distribution, and how it was derived from the observed sentence. What is this interpretation?

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

From the developed equation in 7, we can replace the denominator with the answer we got from 6:

$$\frac{\prod_{w \in V} \theta_{a \rightarrow w}^{n_{a \rightarrow w} + \alpha_{a \rightarrow w} - 1} * \prod_{w \in V} \theta_{b \rightarrow w}^{n_{b \rightarrow w} + \alpha_{b \rightarrow w} - 1} * \prod_{w \in V} \theta_{\times \rightarrow w}^{n_{\times \rightarrow w} + \alpha_{\times \rightarrow w} - 1}}{\frac{1}{48}}$$

Continuing,

$$48 * \theta_{a \rightarrow a} * \theta_{a \rightarrow b} * \theta_{a \rightarrow \times} * \theta_{b \rightarrow a} * \theta_{b \rightarrow b} * \theta_{b \rightarrow \times} = 48 * \theta_{a \rightarrow a} * \theta_{a \rightarrow b}^2 * \theta_{b \rightarrow a} * \theta_{\times \rightarrow a}$$

The "easy" way to interpret the posterior distribution is by considering it to be the probability of getting this sentence divided by the marginal probability of the same sentence.

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

Let's consider the following two sentences: $\times abaaaa \times$ and $\times aaaaba \times$. If we "cut" at the 4th position of both sentences, we get the following:

$$u_1 = aba, x = a, v_1 = aa \text{ and } u_2 = aaa, x = a, v_2 = ba$$

Then, $u_1 x v_2$ is the sentence *abaaba*. This sentence is clearly not a member of L . Therefore, L is not strictly 2-local.

Problem 10: Consider the language $L = \{a^n b^m c^n d^m \mid n, m \in \mathbb{N}\}$, 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 10: Please put your answer in the box below.

Let's consider the following two sentences: $\times aabcccd \times$ and $\times abbccdd \times$. If we "cut" at the 3rd position of both sentences, we get the following:

$u_1 = aa, x = b, v_1 = ccd$ and $u_2 = ab, x = b, v_2 = cdd$

Then, $u_1 x v_2$ is the sentence $aabcdd$. This sentence is clearly not a member of L as the quantity of a symbols does not equal the number of c symbols, likewise for c and d . Therefore, L is not strictly 2-local.

Problem 11: Show that the language $L = \{a^n b^m c^n d^m \mid n, m \in \mathbb{N}\}$ is not strictly k -local, for any value of k .

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

Let's consider the following two sentences: $\times a^{k-1} b^{k-1} c^{k-1} d^{k-1} \times$ and $\times a^{k-1} b^{k-2} c^{k-1} d^{k-2} \times$. If we "cut" at the 3rd position of both sentences, we get the following:

$u_1 = a^{k-1} b^{k-1}, x = c^{k-1}, v_1 = d^{k-1}$ and $u_2 = a^{k-1} b^{k-2}, x = c^{k-1}, v_2 = d^{k-2}$

Then, $u_1 x v_2$ is the sentence $a^{k-1} b^{k-1} c^{k-1} d^{k-2}$. This sentence is clearly not a member of L as the quantity of b symbols does not equal the number of d symbols. Therefore, L is not strictly k -local for any value of k .

Problem 12: In class we proved that $L \in \text{SL}_k \implies k\text{-SSC}(L)$. 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 $L_1 \in \text{SL}_k$ and $L_2 \in \text{SL}_k$, then $L_1 \cap L_2 \in \text{SL}_k$.

Answer 12: Please put your answer in the box below.

Let's say that we have two sentences $s_1 \in L_1$ and $s_2 \in L_2$. This entails $s_1 \in \text{SL}_k$ and $s_2 \in \text{SL}_k$. Furthermore, the intersection of $L_1 \cap L_2$ would include the subset of elements present in the intersection of s_1 and s_2 . Since both $s_1 \in \text{SL}_k$ and $s_2 \in \text{SL}_k$, then their intersection is also in SL_k . Since $L_1 \cap L_2$ is built upon elements from $s_1 \cap s_2 \in \text{SL}_k$, then $L_1 \cap L_2 \in \text{SL}_k$, and k -strictly local languages are closed under intersection.