**Student Name: Erin**

**Collaboration Statement:**

**Total hours spent: 20**

**I discussed ideas with these individuals:**

- **Patrick Feeney**

**I consulted the following resources:**

- **https://www.youtube.com/watch?v=emnfq4txDuI**
- **https://www.youtube.com/watch?v=3OgCcnpZtZ8**

**By submitting this assignment, I affirm this is my own original work that abides by the course collaboration policy.**

**Links: [HW1 instructions] [collab. policy]**

**Contents**

**1a: Problem Statement**

**Let $\rho \in (0.0, 1.0)$ be a Beta-distributed random variable: $p \sim$ Beta$(a, b)$.**

**Show that $\mathbb{E}[\rho] = \frac{a}{a+b}$.**

**Hint: You can use these identities, which hold for all $a > 0$ and $b > 0$:**

$$\Gamma(a) = \int_{t=0}^{\infty} e^{-t} t^{a-1} dt \tag{1}$$

$$\Gamma(a+1) = a\Gamma(a) \tag{2}$$

$$\int_0^1 \rho^{a-1}(1-\rho)^{b-1} d\rho = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \tag{3}$$

**1a: Solution**

**The expected value is the total of all potential decision outcomes multiplied by each possibility's probability:**

$$E(\rho) = \int \rho \cdot f(\rho)\, dx. \tag{4}$$

**The probability density function of the beta distribution is the following:**

$$f(\rho) = Beta(\mu|a, b) = \frac{\Gamma(a+b)}{\Gamma(a) \cdot \Gamma(b)} \cdot \rho^{a-1} \cdot (1-\rho)^{b-1}, 0 \leq x \leq 1 \tag{5}$$

**Then, we combine the expressions (4) and (5) and perform algebraic manipulation:**

$$E(\rho) = \int_0^1 x \cdot \frac{\Gamma(a+b)}{\Gamma(a) \cdot \Gamma(b)} \rho^{a-1} (1-\rho)^{b-1} \, dx \tag{6}$$

$$\tag{7}$$

$$= \int_0^1 \frac{\Gamma(a+b)}{\Gamma(a) \cdot \Gamma(b)} \rho^{(a+1)-1} (1-\rho)^{b-1} \, dx \tag{8}$$

$$\tag{9}$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)} \int_0^1 \frac{\rho^{(a+1)-1} (1-\rho)^{b-1}}{\cdot \Gamma(b)} \, dx \tag{10}$$

$$\tag{11}$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)} \cdot \frac{\Gamma(a+1)}{\Gamma(a+b+1)} \int_0^1 \frac{\Gamma(a+b+1)}{\Gamma(a+1) \cdot \Gamma(b)} \cdot \rho^{(a+1)-1} (1-\rho)^{b-1} \, dx$$

$$\tag{12}$$

**Then, we use the identity** $\Gamma(a+1) = a \cdot \Gamma(a)$ **to the coefficient and simplify it.**

$$= \frac{\Gamma(a+b)}{\Gamma(a)} \cdot \frac{a \cdot \Gamma(a)}{(a+b) \cdot \Gamma(a+b)} \int_0^1 \frac{\Gamma(a+b+1)}{\Gamma(a+1) \cdot \Gamma(b)} \cdot \rho^{(a+1)-1} (1-\rho)^{b-1} \, dx$$

$$\tag{13}$$

$$= \frac{a}{(a+b)} \int_0^1 \frac{\Gamma(a+b+1)}{\Gamma(a+1) \cdot \Gamma(b)} \cdot \rho^{(a+1)-1} (1-\rho)^{b-1} \, dx \tag{14}$$

$$\tag{15}$$

**Recognize that the part in blue is a Beta distribution and substitute** $Beta(\mu|(a+1), b)$ **in it:**

$$= \frac{a}{(a+b)} \int_0^1 Beta(\mu|(a+1), b) \, dx \tag{16}$$

**Since** $\int_0^1 Beta(\mu|(a+1), b) \, dx = 1$**, we have:**

$$= \frac{a}{a+b} \tag{17}$$

**1b: Problem Statement**

**Let $\mu$ be a Dirichlet-distributed random variable: $\mu \sim$ Dir$(a_1, \dots a_V)$.**

**Show that $\mathbb{E}[\mu_w] = \frac{a_w}{\sum_{v=1}^{V} a_v}$, for any integer $w$ that indexes a vocabulary word.**

**Hint: You can use the identity:**

$$\int \mu_1^{a_1-1} \mu_2^{a_2-1} \dots \mu_V^{a_V-1} d\mu = \frac{\prod_{v=1}^{V} \Gamma(a_v)}{\Gamma(a_1 + a_2 \dots + a_V)} \tag{18}$$

**1b: Solution**

$$E[\mu_w] = \int \mu_w \cdot \frac{\Gamma(a_0)}{\Gamma(a_1) \dots \Gamma(a_V)} \cdot \prod_{v=1}^{V} \mu_v^{a_v-1} d\mu \tag{19}$$

$$\tag{20}$$

**Rearrange the terms with $\mu$ in red for easier integration.**

$$= \int \frac{\Gamma(a_0)}{\Gamma(a_1) \dots \Gamma(a_V)} \cdot \mu_w \cdot \mu_w^{a_w-1} \cdot \prod_{\substack{v=1 \\ v \neq w}}^{V} \mu_v^{a_v-1} d\mu \tag{21}$$

$$= \int \frac{\Gamma(a_0)}{\Gamma(a_1) \dots \Gamma(a_V)} \cdot \mu_w^{(a_w+1)-1} \cdot \prod_{\substack{v=1 \\ v \neq w}}^{V} \mu_v^{a_v-1} d\mu \tag{22}$$

$$\tag{23}$$

**Define a set of new parameters:**

$$b_W = a_W + 1 \tag{24}$$

$$b_i = a_i,\ for\ 1 \leq i \leq V \tag{25}$$

**For these parameters, we have these identities:**

$$\Gamma(b_W) = \Gamma(a_W + 1) = a_W \cdot \Gamma(a_W) \tag{26}$$

$$\Gamma(b_i) = \Gamma(a_i) \tag{27}$$

$$b_0 = 1 + \sum_{v=1}^{V} a_i = 1 + a_0 \tag{28}$$

**Combining (26) and (27), we derive the following:**

$$\Gamma(b_1)\dots\Gamma(b_v) = a_w \cdot \Gamma(a_1)\dots\Gamma(a_v) \tag{29}$$

$$\Gamma(a_1)\dots\Gamma(a_v) = \frac{\Gamma(b_1)\dots\Gamma(b_v)}{a_w} \tag{30}$$

$$\tag{31}$$

**Using (28), we derive the following:**

$$\Gamma(b_0) = \Gamma(1 + a_0) \tag{32}$$

$$= a_0 \cdot \Gamma(a_0) \tag{33}$$

$$\Gamma(a_0) = \frac{\Gamma(b_0)}{a_0} \tag{34}$$

$$\tag{35}$$

**Substitute (30) and (31) in the the equation at (22):**

$$= \int \frac{a_w}{a_0} \cdot \frac{\Gamma(b_0)}{\Gamma(b_1)\dots\Gamma(b_V)} \cdot \mu_w^{(a_w+1)-1} \cdot \prod_{\substack{v=1 \\ v\neq w}}^{V} \mu_v^{a_v-1} d\mu \tag{36}$$

$$\tag{37}$$

**Substitute (24) and (24) in the equation at (36):**

$$= \int \frac{a_w}{a_0} \cdot \frac{\Gamma(b_0)}{\Gamma(b_1)\dots\Gamma(b_V)} \cdot \mu_w^{b_w-1} \cdot \prod_{\substack{v=1 \\ v\neq w}}^{V} \mu_v^{b_v-1} d\mu \tag{38}$$

$$\tag{39}$$

**Then, take the coefficients out oand simplify the integral :**

$$= \frac{a_w}{a_0} \cdot \frac{\Gamma(b_0)}{\Gamma(b_1)\dots\Gamma(b_V)} \int \prod_{v=1}^{V} \mu_v^{b_v-1} d\mu \tag{40}$$

$$\tag{41}$$

**Realize that you can rewrite the given identity at (12) in question as:**

$$\int \prod_{v=1}^{V} \mu_v^{b_v-1} d\mu = \int \mu_1^{b1-1} \dots \mu_V^{b_V-1} d\mu \tag{42}$$

$$= \frac{\prod_{v=1}^{V} \Gamma(b_v)}{\Gamma(b_1) + \dots + b_V)} = \frac{\Gamma(b_1) \dots \Gamma(b_V)}{\Gamma(b_0)} \tag{43}$$

**Substitute the simplified version of the give identity at (43) in the integral:**

$$= \frac{a_w}{a_0} \cdot \frac{\Gamma(b_0)}{\Gamma(b_1) \dots \Gamma(b_V)} \cdot \frac{\Gamma(b_1) \dots \Gamma(b_V)}{\Gamma(b_0)} \tag{44}$$

$$\tag{45}$$

$$= \frac{a_w}{a_0} \tag{46}$$

$$= \frac{a_w}{\sum_{v=1}^{V} a_v} \tag{47}$$

**2a: Problem Statement**

**Show that the likelihood of all $N$ observed words can be written as:**

$$p(X_1 = x_1, X_2 = x_2, \dots, X_N = x_N|\mu) = \prod_{v=1}^{V} \mu_v^{n_v} \tag{48}$$

**Hint: It may be helpful to recall the definition of the Categorical PMF using indicator notation:**

$$p(X_n = x_n|\mu) = \prod_{v=1}^{V} \mu_v^{[x_n=v]} \tag{49}$$

**Also, remember the relationship between this bracket notation and the count of how often vocabulary term $v$ appears in the training data: $n_v = \sum_{n=1}^{N}[x_n = v]$**

**2a: Solution**

**Because we assume that each word is conditionally independent of the other words given a parameter vector , the likelihood of all N observed words is:**

$$p(X_1 = x_1, X_2 = x_2, \dots, X_N = x_N|\mu) = \prod_{n=1}^{N} p(X_n = x_n|\mu) \tag{50}$$

**Furthermore, since each individual word is identically distributed according to $\mu$, we can write:**

$$p(X_n = x_n|\mu) = \mu_{x_n} \tag{51}$$

**Then, the likelihood can be written as:**

$$p(X_1 = x_1, X_2 = x_2, \dots, X_N = x_N|\mu) = \prod_{n=1}^{N} \mu_{x_n} \tag{52}$$

**We can reformulate the likelihood as a product over the vocabulary, where each vocabulary probability $\mu_v$ is raised to the power of its corresponding**

**count.**

$$L(\mu) = \prod_{n=1}^{N} \mu_{x_n} = \prod_{v=1}^{V} \mu_v^{n_v} \tag{53}$$

**The Iverson bracket $[x_n = v]$ counts occurrences. It equals 1 if $x_n$ equals $v$ (the $n$-th word is the $v$-th vocabulary term) and 0 otherwise. Therefore, $n_v$, can be expressed using the Iverson bracket as:**

$$n_v = \sum_{n=1}^{N} [x_n = v] \tag{54}$$

### 2b: Problem Statement

**Derive the next-word posterior predictive, after integrating away parameter $\mu$.**

**That is, show that after seeing the $N$ training words, the probability of the next word $X_*$ being vocabulary word $v$ is:**

$$\begin{aligned} p(X_* = v | X_1 = x_1 \dots X_N = x_N) &= \int p(X_* = v, \mu | X_1 = x_1 \dots X_N = x_N) d\mu \\ &= \frac{n_v + \alpha}{N + V\alpha} \end{aligned} \tag{55}$$

**Hint: You will use the expectation of a Dirichlet-distributed random variable that we proved in 1b**

### 2b: Solution

**The posterior predictive for a new observation $X^*$ given the data $D$ is the integral over all possible parameter vectors $\mu$:**

$$p(X^* = v | D) = \int p(X^* = v | \mu, D) p(\mu | D) d\mu \tag{56}$$

**Since $X^*$ is conditionally independent of $D$ given $\mu$, we can simplify the first term in the integral as following:**

$$p(X^* = v | D) = \int p(X^* = v | \mu) p(\mu | D) d\mu \tag{57}$$

**From the problem statement, we know that the posterior $p(\mu | D)$ is Dirichlet with parameters $\alpha + m$. So, we use the equation for the Dirichlet posterior (Eq. 2.41) and substitute $p(\mu | D)$:**

$$p(X^* = v | D) = \int \mu_v \frac{\Gamma(\alpha_0 + N)}{\prod_{k=1}^{K} \Gamma(\alpha_k + m_k)} \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1} d\mu \tag{58}$$

**Recognize that the integral in (57) is the expected value of $\mu_v$ under the Dirichlet distribution, and therefore we don't need to perform an explicit**

**integration:**

$$E[\mu_v] = \frac{\alpha_v + m_v}{\sum_{v=1}^{V}(\alpha_v + m_v)} \tag{59}$$

**We know that, in our Dirichlet distribution, ll $\alpha_k$ are equal to $\alpha$, so $\alpha_v = \alpha$ for all $v$, and $m_v = n_v$, which is the the count of the $v$-th word. So, we substitute these in.**

$$p(X^* = v|D) = E[\mu_v] = \frac{\alpha_v + n_v}{\sum_{v=1}^{V}(\alpha_v + n_v)} \tag{60}$$

**Since $\alpha_v = \alpha$ for all $v$, and $\sum_{v=1}^{V}(\alpha_v + n_v) = V\alpha + N$ because the sum of all $n_v$ is $N$, the total count of words, we have:**

$$p(X^* = v|D) = \frac{n_v + \alpha}{V\alpha + N} \tag{61}$$

**2c: Problem Statement**

**Derive the marginal likelihood of observed training data, after integrating away the parameter $\mu$.**

**That is, show that the marginal probability of the observed $N$ training words has the following closed-form expression:**

$$p(X_1 = x_1 \dots X_N = x_N) = \int p(X_1 = x_1, \dots X_N = x_N, \mu) d\mu \tag{62}$$

$$= \frac{\Gamma(V\alpha) \prod_{v=1}^{V} \Gamma(n_v + \alpha)}{\Gamma(N + V\alpha) \prod_{v=1}^{V} \Gamma(\alpha)} \tag{63}$$

**2c: Solution**

**Using the product rule, we can rewrite the marginal probability of the observed $N$ training words as follows:**

$$p(X_1 = x_1, \dots, X_N = x_N) \tag{64}$$

$$= \int p(X_1 = x_1, \dots X_N = x_N, \mu) d\mu \tag{65}$$

$$= \int p(X_1 = x_1, \dots, X_N = x_N | \mu) p(\mu) d\mu \tag{66}$$

**First, let's evaluate likelihood of the data given the parameter $\mu$. We know that**

$$p(X_1 = x_1, \dots, X_N = x_N | \mu) = \prod_{v=1}^{V} \mu_v^{n_v} \tag{67}$$

**because of conditional independence from 2(a).**

**Second, we evaluate the Dirichlet prior $\rho(\mu)$. Dirichlet prior is**

$$p(\mu) = \frac{1}{B(\alpha)} \prod_{v=1}^{V} \mu_v^{\alpha_v - 1} \tag{68}$$

**where $B(\alpha)$ is the beta function (or the normalization constant for the Dirichlet distribution), and $\alpha_v$ are the parameters of the Dirichlet distribution.**

**Combine the likelihood and the prior:**

$$p(X_1 = x_1, \ldots, X_N = x_N) = \int \left( \prod_{v=1}^{V} \mu_v^{n_v} \right) \left( \frac{1}{B(\alpha)} \prod_{v=1}^{V} \mu_v^{\alpha_v - 1} \right) d\mu \tag{69}$$

$$\tag{70}$$

**Simplify the expression and recognize the new Dirichlet distribution.**

$$= \int \frac{1}{B(\alpha)} \prod_{v=1}^{V} \mu_v^{n_v + \alpha_v - 1} d\mu \tag{71}$$

$$\tag{72}$$

$$= \frac{1}{B(\alpha)} B(n + \alpha) \tag{73}$$

$$\tag{74}$$

**Now, we can use the properties of the gamma function and the definition of the beta function in terms of gamma functions:**

$$B(\alpha) = \frac{\prod_{v=1}^{V} \Gamma(\alpha_v)}{\Gamma\left(\sum_{v=1}^{V} \alpha_v\right)}$$

$$B(n + \alpha) = \frac{\prod_{v=1}^{V} \Gamma(n_v + \alpha_v)}{\Gamma\left(\sum_{v=1}^{V} (n_v + \alpha_v)\right)}$$

**Combining these properties, we get the marginal likelihood:**

$$= \frac{\Gamma\left(\sum_{v=1}^{V} \alpha_v\right)}{\prod_{v=1}^{V} \Gamma(\alpha_v)} \frac{\prod_{v=1}^{V} \Gamma(n_v + \alpha_v)}{\Gamma\left(\sum_{v=1}^{V} (n_v + \alpha_v)\right)} \tag{75}$$

$$\tag{76}$$

$$= \frac{\Gamma(V\alpha) \prod_{v=1}^{V} \Gamma(n_v + \alpha)}{\Gamma(N + V\alpha) \prod_{v=1}^{V} \Gamma(\alpha)} \tag{77}$$