2. EM Alg

a) Consider 5 documents and 3 topics

d_1 has topic 1
d_2 has topic 2
d_3 has topic 3
d_4 has topic 1
d_5 has topic 2.

To maximize the likelihood, we need to maximize the probability

$$p(d_1 = 1, d_2 = 2, d_3 = 3, d_4 = 1, d_5 = 2)$$

$$\Rightarrow p(d_i = 1) = \frac{2}{5} \quad p(d_i = 2) = \frac{2}{5} \quad p(d_i = 3) = 1/5$$

This encourages the topic data we observed.

Thus, $\pi = (\frac{2}{5}, \frac{2}{5}, 1/5)$ is the maximum likelihood estimator.

Let $d_1, \dots, d_n$ be the list of documents,

$$I_k : d_i \longrightarrow \{0,1\} \qquad I_k(d_i) = \begin{cases} 1 & \text{if } d_i \text{ is of topic } k \\ 0 & \text{otherwise} \end{cases}$$

$$\pi_k = \frac{\sum_{i=1}^{n} I_k(d_i)}{n}, \quad n \text{ documents.}$$

Consider documents of topic K

Ex. topic : airplanes

d_1: I like airplanes
d_2: Airplanes are cool
d_3: Airplanes fly.

airplanes $\xrightarrow{\text{encode}}$ [1,0,0,0,0,0]
I $\xrightarrow{\text{encode}}$ [0,1,0,0,0,0]
⋮

Vocab $M = 6$, total words = 8, count(airplanes) = 3

$$\Rightarrow p([1,0,0,0,0,0] \mid \text{topic} = \text{airplanes}) = \frac{3}{8}$$

$$p([0,1,0,0,0,0] \mid \text{topic} = \text{airplanes}) = 1/8 \quad \dots$$

To maximize this situation, given topic K

$$\mu_k(i) = \frac{\text{number of occurrences of the } i^{th} \text{ word in documents of topic } K}{\text{total number of occurrences of all words in docs of topic } K}$$

b) Initialize π, μ_k randomly $\forall k = 1, \dots, K$

Expectation:

$$\varphi_{ki} = p(\mu_k | w_i) = \frac{p(w_i|\mu_k)\, p(\mu_k)}{p(w_i)} = \frac{\prod_{i=1}^{M} (\mu_k(i))^{w(i)} \pi_k}{\sum_{k=1}^{K} \pi_k \left(\prod_{i=1}^{m} \mu_k(i)^{w(i)}\right)}$$

Maximization

π_k = prob of k^{th} topic

Given class k, we have N words per $d_1, \dots, d_n$ n documents, and each word suggests a certain probability of class K

over docs

$$\Rightarrow \pi_k = \frac{\sum_{d_j \in \{d_1, \dots, d_n\}} \sum_{w_i \in d_j} \varphi_{ki}}{n \cdot N}$$

$\mu_k(i)$ = given topic k, prob of drawing i^{th} word

$$= \frac{\sum_{d_j \in \{d_1, \dots, d_n\}} \sum_{w_i \in d_j} \varphi_{ki} \mu_k(i)}{\sum_{d_j \in \{d_1, \dots, d_n\}} \sum_{w_i \in d_j} \mu_k(i)}$$

c) For the previous word, we can have $(M-1)$ possible values, assuming the previous word $\neq$ current word in the M-word vocabulary. π has k parameters (one for each topic) and $\mu_k[i]$ is $\mathbb{R}^m$ for each topic $k \Rightarrow k*M$

In total
$$k \times k*M \times (M-1)$$

More parameters $\Rightarrow$ more complex model. Thus, variance would increase (bias decrease). Compared to the previous model, this model would have a higher variance.