Collapsed Gibbs Sampling (CGS) for Topic Modeling

Corpus Setup

- **Documents:**
  - Document 1: ["apple", "banana", "apple"]
  - Document 2: ["banana", "orange", "orange"]
  - Document 3: ["mango", "grape", "mango", "grape", "lemon"]
- **Vocabulary:** {"apple", "banana", "orange", "grape", "mango", "lemon"}
- **Number of Topics:** 2
- **Hyperparameters:** $\alpha = 0.1$, $\beta = 0.1$

Step 1: Initialization

Each word is randomly assigned a topic (either Topic 0 or Topic 1). The initial topic assignments are:

- Document 1: ["apple" → Topic 0, "banana" → Topic 1, "apple" → Topic 1]
- Document 2: ["banana" → Topic 0, "orange" → Topic 1, "orange" → Topic 0]
- Document 3: ["mango" → Topic 0, "grape" → Topic 1, "mango" → Topic 0, "grape" → Topic 1, "lemon" → Topic 0]

Step 2: Count Matrices (Initial State)

Document-Topic Count Matrix ($n_{d,k}$)

This matrix tracks the number of words in each document assigned to each topic:

$n_{d,k} =$

Document	Topic 0	Topic 1
Doc 1	1	2
Doc 2	2	1
Doc 3	3	2

Topic-Word Count Matrix ($n_{k,w}$)

This matrix tracks how many times each word is assigned to each topic:

$n_{k,w} =$

Topic	apple	banana	orange	grape	mango	lemon
0	1	1	1	0	2	1
1	1	1	1	2	0	0

Topic Count Vector (n_k)

This vector tracks the total number of words assigned to each topic across all documents:

$$n_k = \begin{pmatrix} 6 \\ 5 \end{pmatrix}$$

Step 3: Collapsed Gibbs Sampling (CGS)

We update the topic for each word in the corpus iteratively. Let's continue with Document 3.

Iteration 2: Document 3, Word 2 ("mango")

1. **Remove Current Topic Assignment (Topic 0):** - Decrease counts for "mango" in Topic 0. - Update counts:

$n_{d,k} =$

Document	Topic 0	Topic 1
Doc 1	1	2
Doc 2	2	1
Doc 3	2	2

$n_{k,w} =$

Topic	apple	banana	orange	grape	mango	lemon
0	1	1	1	0	1	1
1	1	1	1	2	0	0

2. **Calculate Conditional Probabilities:**
For Topic 0:

$$P(z = 0|d, w) \propto \left(\frac{n_{d,0} + \alpha}{n_d + K\alpha}\right) \times \left(\frac{n_{0,w} + \beta}{n_0 + V\beta}\right)$$

$$P(z=0) \propto \frac{2+0.1}{4+2\cdot 0.1} \times \frac{1+0.1}{5+6\cdot 0.1} = 0.6818 \times 0.1471 = 0.1002$$

For Topic 1:

$$P(z=1|d,w) \propto \left(\frac{n_{d,1}+\alpha}{n_d+K\alpha}\right) \times \left(\frac{n_{1,w}+\beta}{n_1+V\beta}\right)$$

$$P(z=1) \propto \frac{2+0.1}{4+2\cdot 0.1} \times \frac{0+0.1}{5+6\cdot 0.1} = 0.7273 \times 0.0200 = 0.0145$$

3. **Normalize Probabilities:**

The probabilities are normalized by dividing by the sum of the probabilities for both topics:

$$P(z=0|d,w) = \frac{0.1002}{0.1002+0.0145} = 0.873$$

$$P(z=1|d,w) = \frac{0.0145}{0.1002+0.0145} = 0.127$$

4. **Sample New Topic:**

Based on the probabilities, we sample a new topic for "mango". In this case, we choose Topic 0 with probability 0.873 and Topic 1 with probability 0.127.

Suppose we assign "mango" to Topic 0.

5. **Update Counts:**

After assigning "mango" to Topic 0, the count matrices are updated:

$n_{d,k} =$

Document	Topic 0	Topic 1
Doc 1	1	2
Doc 2	2	1
Doc 3	2	2

$n_{k,w} =$

Topic	apple	banana	orange	grape	mango	lemon
0	1	1	1	0	2	1
1	1	1	1	2	0	0

Conclusion

We have completed the update for the second word in Document 3. The process is repeated for the remaining words in the document. This is how the Collapsed Gibbs Sampling algorithm iterates over each word in the corpus to update topic assignments.