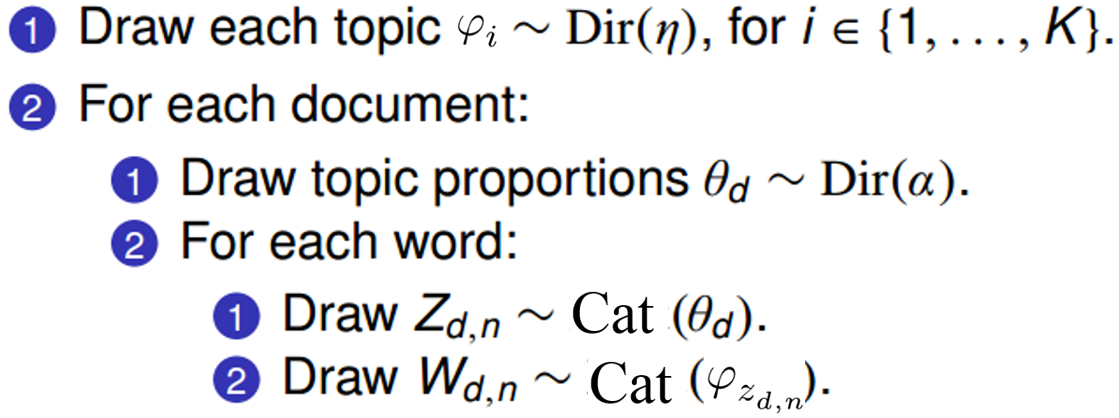
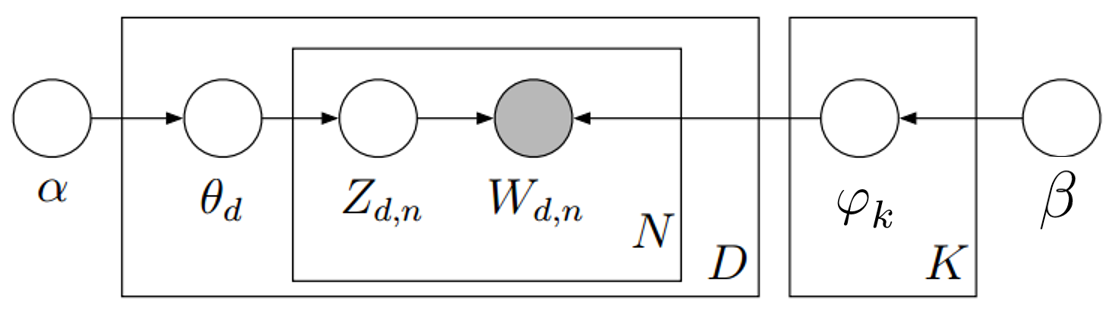
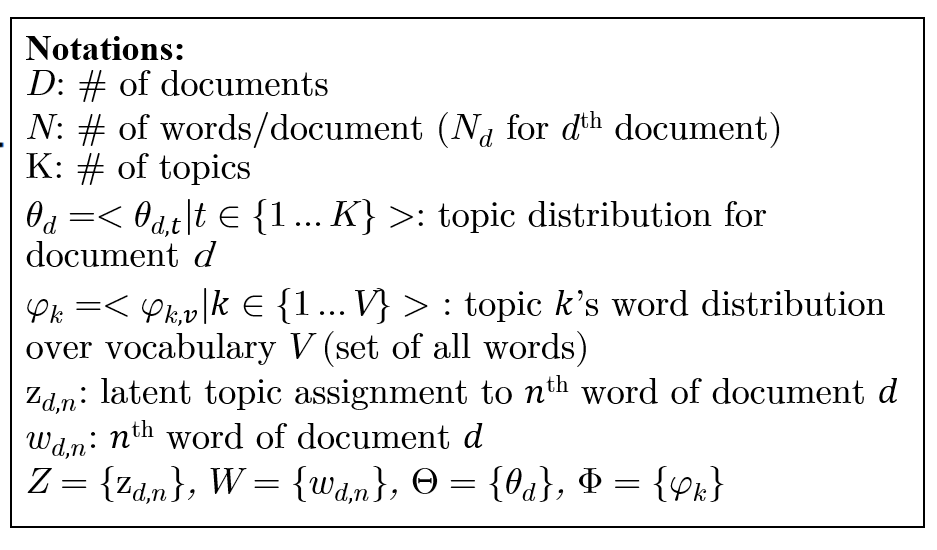
**Gibbs Sampler Derivation for Latent Dirichlet Allocation (Blei et al., 2003)**

**Lecture Notes**

**Arjun Mukherjee (UH)**

1. **Generative process, Plates, Notations**

****

1. **The joint distribution:**

Let us solve for first

From definition,

Now, we know, . Recall that

Recall that has the following PDF:

, where

Also since integrating the PDF over the simplex must equal 1 (from definition), we have

or (Identity A)

Thus,

(1)

In our case, we assume each hyper-parameter of the K dimensional Dir, ,

i.e., the vector

Recall that

(2)

Now since

We have where for a given word, , out of {, …} exactly one of and rest all are zero, i.e., . This follows directly from the categorical distribution because we are sampling a single topic for the word in document . The sampled topic *k* has the probability of .

Thus,

(3)

as where

Now, = # of words in document that were assigned to topic . Let us denote this count by , i.e.,

.

Continuing from (3), we get

(4)

Substituting (4) in (2), we get

(5)

Using (5) and (1), we get

(6)

Since, the document topic distributions, are independent of each other, we can group as follows:

(7)

Using identity A, we get

where

And . Continuing, from (7), we get

(8)

Now Simplifying

Since

(9)

(10)

Since

where for a given word, , in doc out of {, …} exactly one of and rest all are zero, i.e., as we are drawing a word from the topic word distribution, with probability

Continuing from (10) and substituting

Now, = # of times word was assigned to topic . Let this count be denoted by

Thus, we have

(11)

Using (9) and (11), we get

Noting that topic distributions, are independent and grouping the terms

The integral simplifies to where . Thus,

(12)

Thus, using (8) and (12), the full joint distribution of the model can be written as

(13)

1. **The Posterior on and**

**(A)** Computing the posterior distribution for having *observed topic assignments*, in document .

Prior: . Prior probability:

Likelihood: .

Likelihood of this document given : (using (4))

Posterior on using Bayes theorem:

i.e., Posterior on

which is nothing but proportional to the . Here we see conjugate priors in action. Since Dirichlet is the conjugate prior of Categorical distribution, the posterior also takes the form of the prior, i.e., another Dirichlet with added pseudocounts.

In fact one can directly use the properties of conjugate priors to arrive at .

Using the properties of the Dirichlet distribution, one can easily obtain the expected value of the probability mass associated to each topic in the document d as follows:

(14)

**(B)** Computing the posterior distribution for

Using the fact that the Dirichlet is conjugate to the categorical distribution used for word emission process, we can arrive at . Thus,

(15)

1. **Gibbs sampler**

Let be a dimensional random vector, i.e., denotes the collection of all latent topic variables, corresponding to all words in all documents.

Also let us posit a Markov chain over the data and the model, whose stationary distribution converges to the posterior on distribution of .

Also let be denoted by single subscript of ease of notation. For a given token/word, , i.e., the word at document , the Gibbs conditional (sampling distribution) for its latent topic can be constructed as follows:

(16)

where and . Equation (16) gives us the probability that the latent topic variable at is assigned to topic having observed all other topic assignments and words except . Also for subsequent steps the subscript denotes all counts/variables/functional values upon discounting (or not accounting) the token at

Expanding the sampling distribution using (13) we get:

(17)

Thus, the Gibbs sampling distribution for says that it is proportional to the full joint distribution of the model divided by the joint considering the token/word, and its associated topical assignment did not exist in our data/model.

Observing that remains fixed, and corresponds to the topic and words, at some document and some position in that document, we can further simplify the first term in (17) as

(18)

As for all documents other than the numerator and denominator remain exactly the same and cancel out. Now let us see what changes happen in the count vector with or without including the term/word at whose latent topic assignment is . We recall that and # of words in document that were assigned to topic . Hence, we can write

(19)

Expanding (18) using the definition of , we get

Using the result in (19) this further simplifies to (upon canceling out all non terms)

Using the identity , we get

In a similar way, one can also simply the second term in (17) as follows:

Where refers to the vocabulary token which is assigned to . We can thus simplify the full Gibbs conditionals as follows:

The full algorithm for Gibbs sampling is as follows (from Fig 8, [Heinrich, 2008]). There are some minor notation changes which are noted below.

