Week 11 Report

# Nested Chinese restaurant process(nCRP) and hierarchical LDA (hLDA)

Authors: David M. Blei, Thomas L. Griffith

Link:

<https://dl.acm.org/doi/10.1145/1667053.1667056> (main reference)

<https://proceedings.neurips.cc/paper_files/paper/2003/file/7b41bfa5085806dfa24b8c9de0ce567f-Paper.pdf>

The **nCRP** provides a nonparametric prior for the **hLDA** model, replacing the standard Dirichlet distribution used in traditional LDA. This approach allows for a nonparametric model that can organize topics into a hierarchical structure.

## 1. Background

### 1.1 Dirichlet and Beta Distribution:

Dirichlet distribution is a probability distribution on the simplex of nonnegative real numbers that sum to one. U is a random vector.

Beta distribution is a special case of Dirichlet distribution for K = 2 in which the simplex is the unit interval (0,1). In this case, U is a scalar.

### 1.2 Stick-Breaking Construction

Motivation: Consider a collection of nonnegative real numbers,

We want to place a probability distribution of such sequences.

The stick-breaking construction view the interval (0,1) as a unit-length stick. We draw a value V1 from Beta (α1,α2) and break off a fraction V1 form the stick. We let θ1 = V1. The remaining stick will be of length (1 – V1). By continuing the process recursively, we get:

If we set α1 to be 1 α2 to be γ, we obtain a one-parameter stochastic process known as the GEM distribution.

Large values of γ skew the beta distribution towards zero and yield random sequences that assign significant probability t larger integers. Small value of γ yield random sequences that decay more quickly to 0.

Finally, in the following paper, GEM is redefined to a two-parameter variant with the parameter m and π

The standard GEM is the special case when mπ = 1 and γ = (1-m) π.

1.3 Dirichlet Process.

Combines GEM and CRP process 🡺 distribution over distribution

…

## 2. Generative Process

### 2.1 Model Parameters

#### Hyperparameters:

*γ*: Used for nCRP. Controls how likely a customer will sit at a new table.

m: Controls the mean of GEM distribution

π: Controls the concentration of GEM distribution

***η***: A vector used to sample . Controls the sparsity of the topics. Smaller values will lead to topics with most of their probability mass on a small set of words.

#### Variable:

θ: Document-specific topic distribution

β: Topic-specific word distribution. This is the nodes in the tree.

*z*: The topic assignment (level allocation) for each word in the document. The topic is drawn from the multinomial distribution defined by GEM process.

*c*: The path through the tree that a document takes.

*N*: The number of words in the document.

*W*: The size of the vocabulary.

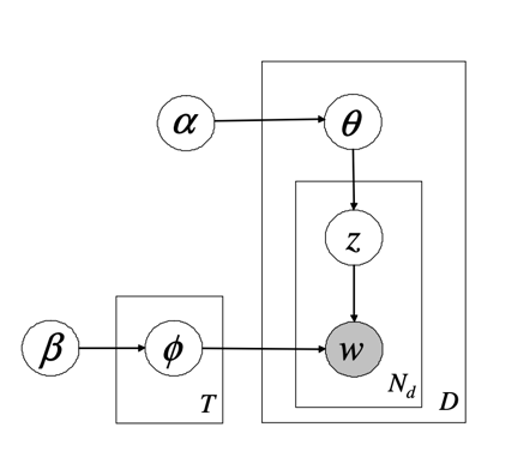
### 2.2 Assumptions

#### 2.2.1 Bag of words assumption

2.2.2 Exchangeability of Documents: Order of document does not matter

2.2.3 Fixed sized tree: The hierarchy is restricted to a fix length of L. However, this can be relaxed.

### 2.3 Outline of Generative Algorithm

A diagram of a diagram

Description automatically generated

Figure (1): (Left) Original LDA. (Right) hLDA.

As we can see from the diagram, the main difference between the 2 model lies in the structure of topic distribution. In hLDA, the L topics are assumed to follow a hierarchical structure where in LDA, the topics has a ‘flat’ structure.

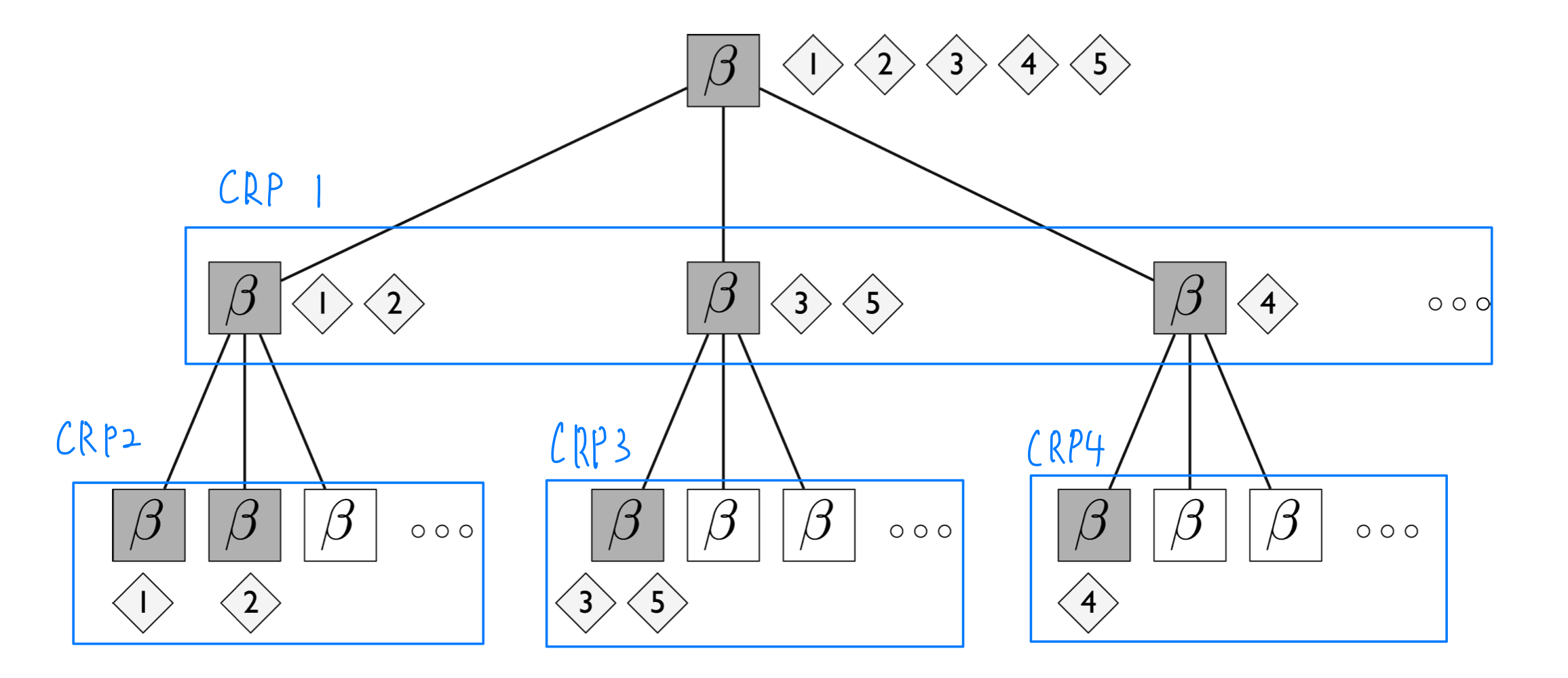


Figure (2): Represents the finite tree derived from the nCRP

|  |
| --- |
| **nCRP outline:** |

For each document:

* 1. Start at the root of the topic tree
  2. Traverse the hierarchy level by level:

2.1 At each level, select a table using the ordinary CRP:

**If no existing table**:

start a new table with Probability = 1.

**Else**:

* + - * 1. Choose an existing table (previously visited topic) with probability proportional to the number of documents already seated there.
        2. Choose a new table (introduce a new topic) with probability proportional to *γ*.

|  |
| --- |
| **hLDA outline:** |

1. For each table in the infinite tree:
   1. Draw a topic *βi ~* Dirichlet (***η***)
2. For each document, d ∈ {1,2,3,4,…,D}
   1. Draw ***cd*** ~ nCRP (*γ)*
   2. Draw a distribution over levels in the tree,
   3. For each word:
      1. Sample a level allocation(topic) *zd,n* base on .
      2. Sample the word from the multinomial word distribution

## 3. Inference Process

The inference processes makes use of Collapsed Gibbs sampling to accesses the 2 variables of interest:

: Topic assignment of the *n*th word in the *d*th document to one of the *L* available topics.

: path assignment for document *m*.

Conceptually, the Gibbs sampler is split into 2 parts. First, given the current path assignment, we sample the level allocation , for word n in document d. Secondly, given the level allocation variables, we sample with the , the path assignment for document m.

### 3.1 Sampling level allocations,

1. The first term is a distribution over level. i.e. what topic should this word take given other topic assignments.
2. The second term is the smoothed frequency of seeing word wd,n allocated to the topic at level zd,n of the path **cd**.

### 3.2 Sampling path,

This expression is an instance of Bayes’s Theorem with as the probability of the data given a particular choice of path and it the prior on paths implied by the nCRP.