Implementation Notes of `ldadf`

Hayato Kobayashi

1 Preliminaries

The following tree structure is assumed as a Dirichlet tree in the implementation of `ldadf`. Note that this implementation is a simplified version of the original LDA-DF, where this one directly encodes the maximal independent sets on a CL-graph into trees, whereas the original one encodes the cliques on each connected component in its complement graph into subtrees.

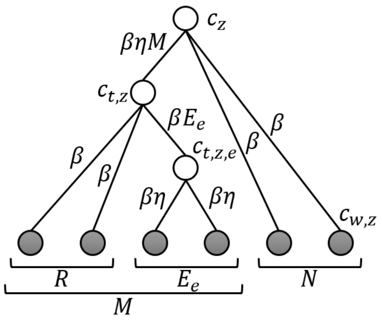

Figure 1: Example of Dirichlet tree.

All variables in the figure are listed in the following table. Each uppercase variable means the number of words corresponding to the leaves in a certain group, and its calligraphic style is used for the set of the words themselves, i.e., $N = |\mathcal{N}|$. Count variables $c_{t,z}$ and $c_{t,z,e}$ are corresponding to the two internal nodes in the above tree and used for inference. The other count variables c_z, $c_{w,z}$, and $c_{d,z}$ are the same as in the standard LDA. See the paper[1] and code[2] for other details.

Table 1: Descriptions of variables.

Var.	Description
N	Number of words in Np
E_e	Number of words in e-th Ep
R	Number of words not in Np/Ep
M	$R + \sum_e E_e$
$c_{t,z}$	Count of topic z for non-Np words in tree t
$c_{t,z,e}$	Count of topic z for words in e-th Ep in tree t
c_z	Count of topic z
$c_{w,z}$	Count of topic z for word w
$c_{d,z}$	Count of topic z for document d

2 Sampling z_i

The sampling equation of topic z_i for collapsed Gibbs sampling is decomposed into the prior part and the likelihood part by using Bayes' theorem, as follows.

$$p(z_i = z \mid \boldsymbol{z}_{-i}, \boldsymbol{q}, \boldsymbol{w}) \propto \\ p(z_i = z \mid \boldsymbol{z}_{-i}, \boldsymbol{q}) p(w_i \mid z_i = z, \boldsymbol{z}_{-i}, \boldsymbol{q}, \boldsymbol{w}_{-i}). \tag{1}$$

The prior part is derived in the same way as LDA.

$$p(z_i = z \mid \boldsymbol{z}_{-i}, \boldsymbol{q}) \propto (c_{d,z} + \alpha). \tag{2}$$

The likelihood part is devided into the following three cases. The first one is for a word in Np, where the word directly falls to the leaf with weight β. The second one is for a word not in Np/Ep, where the word first connects to the internal node with weight $\beta\eta M$ and then falls to the leaf with weight β. The third one is for a word in Ep, where the word passes through the two internal nodes with weights $\beta\eta M$ and βE_e and then falls to the leaf with weight $\beta\eta$.

$$p(w_i \mid z_i = z, \boldsymbol{z}_{-i}, \boldsymbol{q}, \boldsymbol{w}_{-i}) \propto \\ \begin{cases} \dfrac{c_{w,z} + \beta}{c_z + \beta\eta M + \beta N} & \text{(Np} \\ \dfrac{c_{w,z} + \beta}{c_{t,z} + \beta\eta M} \dfrac{c_{t,z} + \beta\eta M}{c_z + \beta\eta M + \beta N} & \text{(o/w} \\ \dfrac{c_{w,z} + \beta\eta}{c_{t,z,e} + \beta\eta E_e} \dfrac{c_{t,z,e} + \beta E_e}{c_{t,z} + \beta M} \dfrac{c_{t,z} + \beta\eta M}{c_z + \beta\eta M + \beta N} & \text{(Ep} \end{cases} \tag{3}$$

3 Sampling q_z

The sampling equation of tree q_z is expressed as the product of the following four rows. The first row represents a prior, which is the weight sum of possible words. This expression is slightly different from the original paper, but it worked well. The remaining three represent a likelihood. The second row represents a probability of a tree from the root node to the internal non-Np node ($c_{t,z}$) and the Np-leaves ($\mathcal{N}$). Similarly, the third row is from the internal non-Np node ($c_{t,z}$) to the internal Ep-nodes ($c_{t,z,e}$) and the normal nodes ($\mathcal{R}$), and the last row is from each Ep-node ($c_{t,z,e}$) to the Ep-leaves ($\mathcal{E}_e$).

$$p(q_z = q \mid \boldsymbol{z}, \boldsymbol{q}_z, \boldsymbol{w}) \propto M\beta \\ \frac{\Gamma(\beta\eta M + \beta N)}{\Gamma(c_z + \beta\eta M + \beta N)} \frac{\Gamma(c_{t,z} + \beta\eta M)}{\Gamma(\beta\eta M)} \prod_w^{\mathcal{N}} \frac{\Gamma(c_{w,z} + \beta)}{\Gamma(\beta)} \\ \frac{\Gamma(\beta M)}{\Gamma(c_{t,z} + \beta M)} \prod_w^{\mathcal{R}} \frac{\Gamma(c_{w,z} + \beta)}{\Gamma(\beta)} \prod_e^{\mathcal{E}} \frac{\Gamma(c_{t,z,e} + \beta E_e)}{\Gamma(\beta E_e)} \\ \prod_e^{\mathcal{E}} \frac{\Gamma(\beta\eta E_e)}{\Gamma(c_{t,z,e} + \beta\eta E_e)} \prod_w^{\mathcal{E}_e} \frac{c_{w,z} + \beta\eta}{\beta\eta}. \tag{4}$$

[1] https://www.aclweb.org/anthology/W11-3905
[2] https://github.com/hakobayato/ldadf