Sampling Table Configurations for the Hierarchical Poisson-Dirichlet Process

Changyou Chen[1,2], Lan Du[1,2], and Wray Buntine[2,1]

[1]Research School of Computer Science,
The Australian National University,
Canberra, ACT, Australia
[2]National ICT, Canberra, ACT, Australia
{Changyou.Chen,Lan.Du,Wray.Buntine}@nicta.com.au

This is a more detailed description for sampling the HDP-LDA model in the paper (eq.12 – 14), containing errata of eq.13 in the paper (missing a term containing b_0 and b_1).

1. If $\forall j', t'_{j'k} = 0$, there is only one possible seating: create a new table in restaurant $j > 0$ and then create a new table at $j = 0$, *e.g.*, $u_l = 0$:

$$P_r(z_l = k_{new}, u_l = 0 \,|\, \boldsymbol{z}_{1:J} - z_l, \boldsymbol{u}_{1:J} - u_l) \propto \frac{b_0 b_1}{b_0 + \sum_k Tt[k]} \frac{\gamma_l + M_{kl}}{\sum_{l'}(\gamma_{l'} + M_{kl'})} \tag{1}$$

2. If $t'_{jk} \neq 0, t'_{0k} \neq 0$, there are two possibilities: 1) create a new table at $j > 0$, thus $u_l = 1$ and $t''_{jk} \neq t'_{jk}$; 2) sit on an existing table, thus $u_l = 2$ (meaning no table created) and $t''_{jk} = t'_{jk}$:

$$\begin{aligned} &P_r(z_l = k, u_l = u \,|\, \boldsymbol{z}_{1:J} - z_l, \boldsymbol{u}_{1:J} - u_l) \qquad (2) \\ &\propto \left(\frac{b_1}{b_0}\right)^{t''_{jk} \neq t'_{jk}} \frac{S^{n''_{jk}}_{t''_{jk},0}}{S^{n'_{jk}}_{t'_{jk},0}} \frac{\left(t''_{jk}\right)^{\delta_{t''_{jk} \neq t'_{jk}}} \left(n''_{jk} - t''_{jk}\right)^{\delta_{n''_{jk} - t''_{jk} \neq n'_{jk} - t'_{jk}}}}{\left(n''_{jk}\right)^{\delta_{n''_{jk} \neq n'_{jk}}}} \frac{\gamma_l + M_{kl}}{\sum_{l'}(\gamma_{l'} + M_{kl'})} \end{aligned}$$

3. If $t'_{jk} = 0, t'_{0k} \neq 0$, there is only one possibility, which is to create a new table at $j > 0$ ($u_l = 1$), but can not create a new table at $j = 0$ because t_{0k} is at most 1 due to the property of the DP:

$$\begin{aligned} &P_r(z_l = k, u_l = 1 \,|\, \boldsymbol{z}_{1:J} - z_l, \boldsymbol{u}_{1:J} - u_l) \\ &\propto \frac{b_1 Tt[k]^2}{(Tt[k] + 1)(\sum_k Tt[k] + b_0)} \frac{\gamma_l + M_{kl}}{\sum_{l'}(\gamma_{l'} + M_{kl'})} \end{aligned} \tag{3}$$

where $Tt[k]$ denotes the number of tables serving dish k (*i.e.*, topic k), M_{kl} indicates the total number of words l assigned to k in the document collection.