Parallel Markov Chain Monte Carlo for Pitman-Yor Mixture Models: Supplement

1 Appendix

1.1 Theorem

Theorem 1 (Auxiliary variable representation for the PYMM)**.** *For $\alpha \geq 0$ we can re-write the generative process for a PYMM as*

$$D_j \sim \text{PY}\left(d, \frac{\alpha}{P}, H\right), \quad \phi \sim \text{Dirichlet}\left(\frac{\alpha}{P}, \dots, \frac{\alpha}{P}\right), \\ \pi_i \sim \phi, \qquad \theta_i \sim D_{\pi_i}, \qquad x_i \sim f(\theta_i), \tag{1}$$

for $j = 1, \dots, P$ and $i = 1, \dots, N$. The posterior distribution over the θ_i remains the same as

$$D \sim PY(\alpha, d, H), \qquad \theta_i \sim D, \qquad x_i \sim f(\theta_i).$$

.

Proof. We will prove that the posterior predictive will be the same as that of Pitman-Yor Mixture Model.

Let $\theta_1, \theta_2, \dots$ be a sequence of random variable distributed according to $G \sim PY(d, \alpha, H)$. Then the conditional distribution of θ_{n+1} given $\theta_1, \dots, \theta_n$ where G has been integrated is given by

$$\begin{aligned} \theta_{n+1}|\theta_1, \dots, \theta_n \quad \sim \quad & \sum_{l=1}^{n} \frac{1}{n+\alpha} \delta_{\theta_l}(\theta_{n+1}) + \frac{\alpha}{n+\alpha} H \\ & - \frac{d}{n+\alpha} \delta(\{\sum_{i=1}^{n} \delta_{\theta_l}(\theta_{n+1})\} \geq 1) \\ & + \frac{d}{n+\alpha} (\sum_{unique\theta_i} 1) H \,. \end{aligned} \tag{2}$$

For the model following theorem 1 the conditional distribution of θ_{n+1} given $\theta_1, \dots, \theta_n$ where D_j $\forall j$ and ϕ has been integrated out is (N_j is the number of points on processor j):

$$\begin{aligned} & \theta_{n+1}|\theta_1, \dots, \theta_n \\ & \quad \sim \sum_{j=1}^{P} P(\pi_{n+1} = j|\pi_1, \dots, \pi_n) \\ & \qquad \cdot P(\theta_{n+1}|\pi_{n+1} = j, \pi_1, \dots, \pi_n, \theta_1, \dots \theta_n, H) \\ & \quad = \sum_{j} \frac{N_j + \alpha/P}{n+\alpha} \\ & \qquad \left\{ \sum_{l=1}^{n} \frac{1}{N_j + \alpha/P} \delta_{\theta_l}(\theta_{n+1}) \delta_j(\pi_i) \right. \\ & \qquad - \frac{d}{N_j + \alpha/P} \delta(\{\sum_{i=1}^{n} \delta_{\theta_l}(\theta_{n+1}) \delta_j(\pi_i)\} \geq 1) \\ & \qquad + \frac{\alpha/P}{N_j + \alpha/P} H \\ & \qquad \left. + \frac{d}{N_j + \alpha/P} (\sum_{unique\theta_i} \delta_j(\pi_i)) H \right\} \\ & \quad = \sum_{l=1}^{n} \frac{1}{n+\alpha} \delta_{\theta_l}(\theta_{n+1}) + \frac{\alpha}{n+\alpha} H \\ & \qquad - \frac{d}{n+\alpha} \delta(\{\sum_{i=1}^{n} \delta_{\theta_l}(\theta_{n+1})\} \geq 1) \\ & \qquad + \frac{d}{n+\alpha} (\sum_{unique\theta_i} 1) H \,. \end{aligned} \tag{3}$$

□

1.2 Metropolis Hastings acceptance probabilities

We just need the likelihood ratio to calculate MH acceptance probabilities since $q(\{\pi_i\} \rightarrow \{\pi_i^*\}) = q(\{\pi_i^*\} \rightarrow \{\pi_i\})$

1.2.1 PYMM

For the Pitman-Yor mixture model the likelihood ratio is given by:

$$\begin{aligned}
&\frac{p(\{\pi_i^*\})}{p(\{\pi_i\})} \\
&= \frac{p(\{x_i\}|\pi_i^*)p(\{\pi_i^*\}|\alpha, P)}{p(\{x_i\}|\pi_i)p(\{\pi_i\}|\alpha, P)} \\
&= \frac{p(\{z_i\}|\pi_i^*)p(\{\pi_i^*\}|\alpha, P)}{p(\{z_i\}|\pi_i)p(\{\pi_i\}|\alpha, P)} \\
&= \prod_{j=1}^{P} \frac{\Gamma(N_j^* + \alpha/P)}{\Gamma(N_j + \alpha/P)} \frac{(\alpha/P)^{(d;K_j^*-1)}}{(\alpha/P)^{(d;K_j-1)}} \\
&\quad \frac{(\alpha/P + 1 - d)^{(1;N_j-1)}}{(\alpha/P + 1 - d)^{(1;N_j^*-1)}} \\
&\quad \prod_{i=1}^{\max(N_j, N_j^*)} [(1-d)^{(1;i-1)})]^{(a_{ij}^* - a_{ij})} \frac{a_{ij}!}{a_{ij}^*!}
\end{aligned} \tag{4}$$

where

$$(a)^{(b;c)} = \begin{cases} 1 & \text{if } c = 0 \\ a(a+b)\dots(a+(c-1)b) & \text{for } c = 1, 2, \dots \end{cases}$$

Proof. Let N_j be the number of points on processor j and n_{jk} be the number of points in cluster k on processor j. Let K_j be the total number of cluster on processor j and a_{ij} is the number of cluster of size i on cluster j. The probability of the processor allocations is described by the Dirichlet compound multinomial, or multivariate Pólya, distribution,

$$\begin{aligned}
p(\{\pi_i\}|\alpha, P) =& \frac{N!}{\prod_{j=1}^{P} N_j!} \frac{\Gamma(\sum_{j=1}^{P} \alpha/P)}{\Gamma(N + \sum_{j=1}^{P} \alpha/P)} \\
&\cdot \prod_{j=1}^{P} \frac{\Gamma(N_j + \alpha/P)}{\Gamma(\alpha/P)} \\
=& \frac{N!}{\prod_{j=1}^{P} N_j!} \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)} \prod_{j=1}^{P} \frac{\Gamma(N_j + \alpha/P)}{\Gamma(\alpha/P)},
\end{aligned}$$

where $N = \sum_{j=1}^{P} N_j$ is the total number of data points. So,

$$\frac{p(\{\pi_i^*\}|\alpha, P)}{p(\{\pi_i\}|\alpha, P)} = \prod_{j=1}^{P} \frac{N_j!}{N_j^*!} \frac{\Gamma(N_j^* + \alpha/P)}{\Gamma(N_j + \alpha/P)}.$$

Conditioned on the processor indicators, the probability of the data can be written

$$p(\{z_i\}|\{\pi_i\}) = \prod_{j=1}^{P} p(\{n_{jk}\}|N_j),$$

where n_{jk} is the number of data points in the kth on processor j. This can be found by two parameter generalization of Ewens random partition structure [TODO cite Pitman].

$$\begin{aligned}
p(\{n_{jk}\}|N_j) =& \frac{N_j!}{\prod_{k=1}^{K_j} n_{jk}!} \frac{(\alpha/P)^{(d;K_j-1)}}{(\alpha/P + 1 - d)^{(1;N_j-1)}} \\
&\prod_{i=1}^{N_j} \frac{[(1-d)^{(1:i-1)}]^{a_{ij}}}{a_{ij}!}
\end{aligned}$$

Thus

$$\begin{aligned}
\frac{P(\{n_{jk}^*\}|\alpha, \pi)}{P(\{n_{jk}\}|\alpha, \pi)} =& \prod_{j=1}^{P} \frac{N_j^*!}{N_j!} \frac{(\alpha/P)^{(d;K_j^*-1)}}{(\alpha/P)^{(d;K_j-1)}} \\
&\frac{(\alpha/P + 1 - d)^{(1;N_j-1)}}{(\alpha/P + 1 - d)^{(1;N_j^*-1)}} \\
&\prod_{i=1}^{\max(N_j, N_j^*)} [(1-d)^{1;i-1})]^{(a_{ij}^* - a_{ij})} \frac{a_{ij}!}{a_{ij}^*!}
\end{aligned}$$

Thus giving equation 4 □

Specific case

Let us assume that we want to calculate the acceptance probability of transferring cluster k_1 of size $i_1 = n_{j_1 k_1}$ from processor j_1 to cluster j_2. Assume that $(N_{j_1} - n_{j_1 k_1}) > 0$ (ie there is atleast one point in the processor j_1 after removing cluster k_1), $K_{j1} > 1$ (its the same as the assumption before since if there is atleast one point in processor j_1 other than in cluster k_1 then it has more than one cluster) and $K_{j2} > 0$ (atleast one cluster in processor j_2) Then the transfer probability is given by

$$\begin{aligned}
&\frac{p(\{\pi_i^*\})}{p(\{\pi_i\})} \\
&= \frac{\Gamma(N_{j_1} - n_{j_1 k_1} + \alpha/P)\,\Gamma(N_{j_2} + n_{j_1 k_1} + \alpha/P)}{\Gamma(N_{j_1} + \alpha/P)\,\Gamma(N_{j_2} + \alpha/P)} \\
&\frac{\Gamma(N_{j_1} + \alpha/P - d)\,\Gamma(N_{j_2} + \alpha/P - d)}{\Gamma(N_{j_1} - n_{j_1 k_1} + \alpha/P - d)\,\Gamma(N_{j_2} + n_{j_1 k_1} + \alpha/P - d)} \\
&\frac{\alpha/P + (K_{j2} - 1)d}{\alpha/P + (K_{j1} - 2)d} \quad \frac{a_{i_1 j_1}}{a_{i_1 j_2} + 1}
\end{aligned} \tag{5}$$

When $d = 0$ PYMM is same as DPMM. The acceptance ratio for DPMM is $\frac{a_{i_1 j_1}}{a_{i_1 j_2}+1}$ which can also be obtained by setting $d = 0$ in equation 5.

Proof. If $c > 0$ then

$$(a)^{(1;c)} = \frac{\Gamma(a+c)}{\Gamma(a)}$$

Also $a^*_{i_1 j_1} = a_{i_1 j_1} - 1$, $a^*_{i_1 j_2} = a_{i_2 j_2} + 1$, $N^*_{j_1} = N^*_{j_1} - n_{j_1 k_1}$ and $N^*_{j_2} = N_{j_2} - n_{j_1 k_1}$. Substitute these in 4 and cancel terms to get 5 □

1.2.2 Hierarchical Version

For the hierarchical auxiliary variable PY model discussed in **??** the ratio is given by

$$\frac{p(\{x_{mi}\}|\{\pi^*_{mi}\gamma, \boldsymbol{\xi}^*, \alpha, P)}{p(\{x_{mi}\}|\{\pi_{mi}\gamma, \boldsymbol{\xi}, \alpha, P)} \frac{p(\{\pi^*_{mi}\}|\gamma, \boldsymbol{\xi}^*)}{p(\{\pi_{mi}\}|\gamma, \boldsymbol{\xi}^)} \frac{p(\boldsymbol{\xi}^*|\alpha, P)}{p(\boldsymbol{\xi}|\alpha, P)}. \tag{6}$$

We consider an equivalent Chinese restaurant franchise representation [**?**], where each data point is associated with a table (corresponding to clustering in the lower-level DP), and each table is associated with a dish (corresponding to clustering in the upper-level PY).

Let $\mathbf{t}_j$ be the count vector for the top-level PY on processor j – in Chinese restaurant franchise terms, t_{jd} is the number of tables on processor j serving dish d. Let $\mathbf{n}_{jm}$ be the count vector for the mth bottom-level DP on processor j – in Chinese restaurant franchise terms, n_{jmk} is the number of customers in the mth restaurant sat at the kth table of the jth processor. Let T_{mj} be the total number of occupied tables from the mth restaurant on processor j, and let U_j be the total number of unique dishes on processor j. Let a_{jmi} be the total number of tables in restaurant m on processor j with exactly i customers, and b_{ji} be the total number of dishes on processor j served at exactly i tables. We use the notation $n_{jm\cdot} = \sum_k n_{jmk}$, $T_{\cdot j} = \sum_m T_{mj}$, etc.

Since the Metropolis-Hastings step does not change the table and dish assignments of the data, the likelihood ratio in Eq. 6 can be re-written as:

$$\frac{p(\{t^*_{jd}\}, \{n^*_{jmk}\}|\{\pi^*_{mi}\gamma, \boldsymbol{\xi}^*, \alpha, P)}{p(\{t_{jd}\}, \{n_{jmk}\}|\{\pi_{mi}\gamma, \boldsymbol{\xi}, \alpha, P)} \cdot \frac{p(\{\pi^*_{mi}\}|\gamma, \boldsymbol{\xi}^*)}{p(\{\pi_{mi}\}|\gamma, \boldsymbol{\xi}^)} \frac{p(\boldsymbol{\xi}^*|\alpha, P)}{p(\boldsymbol{\xi}|\alpha, P)}. \tag{7}$$

The first term in the Eq. 7 is the ratio of the joint probabilities of the topic- and table-allocations in the local HDPs. This can be obtained by applying the Ewen's sampling formula to both top-level PY and bottom-level DPs.

$$p(\{n_{jmk}\}|\gamma, \boldsymbol{\xi}) = \prod_{m=1}^{M}\prod_{j=1}^{P}(\gamma\xi_j)^{T_{mj}} \frac{n_{jm\cdot}!}{\prod_{k=1}^{T_{mj}} n_{jmk}!} \frac{\Gamma(\gamma\xi_j)}{\Gamma(\gamma\xi_j + n_{jm\cdot})} \prod_{i=1}^{N_j} \frac{1}{a_{jmi}!},$$

and

$$p(\{t_{jd}\}|\alpha, P) = \prod_{j=1}^{P} \frac{T_{\cdot j}!}{\prod_{d=1}^{U_j} t_{jd}!} \frac{(\alpha/p)^{(d;U_j-1)}}{(\alpha/P+1-d)^{(1;T_j-1)}} \prod_{i=1}^{T_{\cdot j}} \frac{[(1-d)^{(1:i-1)}]^{b_{ji}}}{b_{ji}},$$

so

$$\begin{aligned}\frac{p(\{t^*_{jd}\}, \{n^*_{jmk}\}|\{\pi^*_{mi}\gamma, \boldsymbol{\xi}^*, \alpha, P))}{p(\{t_{jd}\}, \{n_{jmk}\}|\{\pi_{mi}\gamma, \boldsymbol{\xi}, \alpha, P))} &= \prod_{j=1}^{P} \frac{(\xi^*_j)^{T^*_{\cdot j}}}{(\xi_j)^{T_{\cdot j}}} \frac{T^*_{\cdot j}!}{T_{\cdot j}!} \frac{(\alpha/P)^{(d;U^*_j-1)}}{(\alpha/P)^{(d;U_j-1)}} \left(\frac{\Gamma(\gamma\xi^*_j)}{\Gamma(\gamma\xi_j)}\right)^M \\ &\frac{(\alpha/P+1-d)^{(1;T_j-1)}}{(\alpha/P+1-d)^{(1;T^*_j-1)}} \\ &\cdot \left\{\prod_{i=1}^{\max(T_{\cdot j}, T^*_{\cdot j})} [(1-d)^{(1;i-1)}]^{b^*_{ji}-b_{ji}} \frac{b_{ji}!}{b^*_{ji}!}\right\} \\ &\prod_{m=1}^{M} \frac{n^*_{jm\cdot}!}{n_{jm\cdot}!} \frac{\Gamma(\gamma\xi_j + n_{jm\cdot})}{\Gamma(\gamma\xi^*_j + n^*_{jm\cdot})} \prod_{i=1}^{\max(N_j, N^*_j)} \frac{a_{jmi}!}{a^*_{jmi}!}.\end{aligned} \tag{8}$$

The probability of the processor assignments is given by:

$$p(\{\pi_{mi}\}|\gamma, \boldsymbol{\xi}) = \prod_{m=1}^{M} \frac{n_{\cdot m\cdot}!}{\prod_{j=1}^{P} n_{jm\cdot}!} \frac{\Gamma(\gamma)}{\Gamma(n_{\cdot m\cdot} + \gamma)} \prod_{j=1}^{P} \frac{\Gamma(\gamma\xi_j + n_{jm\cdot})}{\Gamma(\gamma\xi_j)},$$

so the second term is given by

$$\frac{p(\{\pi^*_{mi}\}|\gamma, \boldsymbol{\xi}^*)}{p(\{\pi_{mi}\}|\gamma, \boldsymbol{\xi})} = \prod_{j=1}^{P} \left(\frac{\Gamma(\gamma\xi_j)}{\Gamma(\gamma\xi^*_j)}\right)^M \prod_{m=1}^{M} \frac{n_{jm\cdot}!}{n^*_{jm\cdot}!} \frac{\Gamma(\gamma\xi^*_j + n^*_{jm\cdot})}{\Gamma(\gamma\xi_j + n_{jm\cdot})}. \tag{9}$$

The third term is given by

$$\frac{p(\boldsymbol{\xi}^*|\alpha, P)}{p(\boldsymbol{\xi}|\alpha, P)} = \prod_{j=1}^{P} \left(\frac{\xi^*_j}{\xi_j}\right)^{\frac{\alpha}{P}}. \tag{10}$$

Combining the ratio is given by

$$\begin{aligned} r = &\prod_{j=1}^{P} \frac{(\xi^*_j)^{(T^*_{\cdot j}+\alpha/P)}}{((\xi_j)^{(T_{\cdot j}+\alpha/P)}} \frac{T^*_{\cdot j}!}{T^*_{\cdot j}!} \frac{(\alpha/P)^{(d;U^*_j-1)}}{(\alpha/P)^{(d;U_j-1)}} \\ &\frac{(\alpha/P+1-d)^{(1;T_{\cdot j}-1)}}{(\alpha/P+1-d)^{(1;T^*_{\cdot j}-1)}} \\ &\cdot \left\{\prod_{i=1}^{\max(T_{\cdot j}, T^*_{\cdot j})} [(1-d)^{(1;i-1)}]^{b^*_{ji}-b_{ji}} \frac{b_{ji}!}{b^*_{ji}!}\right\} \\ &\prod_{m=1}^{M} \prod_{i=1}^{\max(N_j, N^*_j)} \frac{a_{jmi}!}{a^*_{jmi}!}.\end{aligned} \tag{11}$$