Supplemental Material for Parallel Sampling of DP Mixture Models using Sub-Clusters Splits

Jason Chang
CSAIL, MIT
jchang7@csail.mit.edu

John W. Fisher III
CSAIL, MIT
fisher@csail.mit.edu

A Posterior Distribution of Weights

In this section, we show the derivation of the posterior distribution over cluster-weights, π, conditioned on the cluster labels, z. We begin with the definition of a Dirichlet process from [1].

Definition A.1 (Dirichlet Process)**.** *Let H be a measure on a measureable space, Ω. If for any finite partition, $(A_1, A_2, \cdots, A_K)$ of the space, the measure, G, on the partition follows the following Dirichlet distribution*

$$(G(A_1), G(A_2), \cdots, G(A_K)) \sim \text{Dir}(\alpha H(A_1), \alpha H(A_2), \cdots, \alpha H(A_K)), \tag{1}$$

for some positive scalar α, then G is said to be a Dirichlet process with concentration parameter α and base measure H.

For each unique cluster label, k, there is a corresponding parameter θ_k, drawn from the base measure. Thus, the posterior base measure is the sum of H with delta functions located at each θ_k, each with area N_k. Here, N_k denotes the number of data points assigned to cluster k. Conditioned on the K unique clusters, we can then form a partition of the space as follows:

$$\left\{A_1 = \delta_{\theta_1}, A_2 = \delta_{\theta_2}, \cdots, A_K = \delta_{\theta_K}, A_{\tilde{K}+1} = \Omega \setminus (\delta_{A_1} \cup \delta_{A_2} \cup \cdots \cup \delta_{A_K})\right\}. \tag{2}$$

Using the definition of Dirichlet processes, this partition has the following distribution

$$(\pi_1, \pi_2, \cdots, \pi_K, \tilde{\pi}_{K+1}) \sim \text{Dir}(N_1, N_2, \cdots, N_K, \alpha). \tag{3}$$

B Auxiliary Variable Distributions

For the naive choice for auxiliary parameter distributions of Equations 11-13, reproduced here

$$p(\overline{\pi}_k) = \text{Dir}(\overline{\pi}_{k,\ell}, \overline{\pi}_{k,r}; \alpha, \alpha), \tag{4}$$

$$p(\overline{\theta}_k) = f_\theta(\overline{\theta}_{k,\ell}; \lambda) f_\theta(\overline{\theta}_{k,r}; \lambda), \tag{5}$$

$$p(\overline{z}|\overline{\pi}, \theta, x, z) = \prod_k \prod_{\{i; z_i = k\}} \frac{\overline{\pi}_{k,\overline{z}_i} f_x(x_i; \overline{\theta}_{k,\overline{z}_i})}{\overline{\pi}_{k,\ell} f_x(x_i; \overline{\theta}_{k,\ell}) + \overline{\pi}_{k,r} f_x(x_i; \overline{\theta}_{k,r})}, \tag{6}$$

the joint distribution for auxiliary parameters for cluster k can be expressed as

$$\begin{aligned} &p(\overline{\pi}_k, \overline{\theta}_k, \overline{z}_{\{k\}} | x, \pi, z, \theta) \\ &= \text{Dir}(\overline{\pi}_{k,\ell}, \overline{\pi}_{k,r}; \alpha, \alpha) f_\theta(\overline{\theta}_{k,\ell}; \lambda) f_\theta(\overline{\theta}_{k,r}; \lambda) \prod_{\{i; z_i = k\}} \frac{\overline{\pi}_{k,\overline{z}_i} f_x(x_i; \overline{\theta}_{k,\overline{z}_i})}{\overline{\pi}_{k,\ell} f_x(x_i; \overline{\theta}_{k,\ell}) + \overline{\pi}_{k,r} f_x(x_i; \overline{\theta}_{k,r})} \qquad (7) \\ &= \text{Dir}(\overline{\pi}_{k,\ell}, \overline{\pi}_{k,r}; \alpha, \alpha) C(x, z, \overline{\pi}_k, \overline{\theta}_k) \prod_{s=\{\ell,r\}} \overline{\pi}_{k,s}^{N_{k,s}} f_x(x_{\{k,s\}}; \overline{\theta}_{k,s}) f_\theta(\overline{\theta}_{k,s}; \lambda). \qquad (8) \end{aligned}$$

where the function C is defined to be

$$C(x, z, \overline{\pi}_k, \overline{\theta}_k) \triangleq \prod_{\{i;z_i=k\}} \frac{1}{\overline{\pi}_{k,\ell} f_x(x_i; \overline{\theta}_{k,\ell}) + \overline{\pi}_{k,r} f_x(x_i; \overline{\theta}_{k,r})}. \tag{9}$$

By simply ignoring terms that do not correspond to the variable of interest, the posterior distributions for the the sub-cluster weights and parameters can be expressed as

$$\begin{aligned} p(\overline{\pi}_k|\bullet) &\propto C(x, z, \overline{\pi}_k, \overline{\theta}_k) \operatorname{Dir}(\overline{\pi}_{k,\ell}, \overline{\pi}_{k,r}; \alpha, \alpha) \, \overline{\pi}_{k,\ell}^{N_{k,\ell}} \, \overline{\pi}_{k,r}^{N_{k,r}} && (10)\\ &= C(x, z, \overline{\pi}_k, \overline{\theta}_k) \operatorname{Dir}(\overline{\pi}_{k,\ell}, \overline{\pi}_{k,r}; \alpha + N_{k,\ell}, \alpha + N_{k,r}), && (11)\\ p(\overline{\theta}_k|\bullet) &\propto C(x, z, \overline{\pi}_k, \overline{\theta}_k) \, f_\theta(\overline{\theta}_{k,\ell}; \lambda) \, f_\theta(\overline{\theta}_{k,r}; \lambda) \, f_x(x_{\{k,\ell\}}; \theta_{k,\ell}) \, f_x(x_{\{k,r\}}; \theta_{k,r}) && (12)\\ &\propto C(x, z, \overline{\pi}_k, \overline{\theta}_k) \, f_\theta(\overline{\theta}_{k,\ell}; \lambda^*_{k,\ell}) \, f_\theta(\overline{\theta}_{k,r}; \lambda^*_{k,r}), && (13) \end{aligned}$$

where $\bullet$ is used to denote all other variables, and we have assumed conjugate priors for explanation purposes. The term, $C(x, z, \overline{\pi}_k, \overline{\theta}_k)$, complicates these distributions because they no longer follow the form of regular-cluster parameters.

If the sub-cluster parameters follow the distribution of Equation 14 instead of Equation 12, reproduced here,

$$\begin{aligned} p(\overline{\theta}_k|x, z, \pi) &\propto f_\theta(\overline{\theta}_{k,\ell}; \lambda) f_\theta(\overline{\theta}_{k,r}; \lambda) \textstyle\prod_{\{i;z_i=k\}} \left(\overline{\pi}_{k,\ell} f_x(x_i; \overline{\theta}_{k,\ell}) + \overline{\pi}_{k,r} f_x(x_i; \overline{\theta}_{k,r})\right) && (14)\\ &= f_\theta(\overline{\theta}_{k,\ell}; \lambda) f_\theta(\overline{\theta}_{k,r}; \lambda) \, C(x, z, \overline{\pi}_i, \overline{\theta}_k)^{-1}, && (15) \end{aligned}$$

the joint distribution can be expressed as

$$\begin{aligned} &p(\overline{\pi}_k, \overline{\theta}_k, \overline{z}_{\{k\}}|x, \pi, z, \theta)\\ &= \operatorname{Dir}(\overline{\pi}_{k,\ell}, \overline{\pi}_{k,r}; \alpha, \alpha) f_\theta(\overline{\theta}_{k,\ell}; \lambda) f_\theta(\overline{\theta}_{k,r}; \lambda) \textstyle\prod_{\{i;z_i=k\}} \overline{\pi}_{k,\overline{z}_i} f_x(x_i; \overline{\theta}_{k,\overline{z}_i}) && (16)\\ &= \operatorname{Dir}(\overline{\pi}_{k,\ell}, \overline{\pi}_{k,r}; \alpha, \alpha) \textstyle\prod_{s=\{\ell,r\}} \overline{\pi}_{k,s}^{N_{k,s}} f_x(x_{\{k,s\}}; \overline{\theta}_{k,s}) f_\theta(\overline{\theta}_{k,s}; \lambda). && (17) \end{aligned}$$

Following a similar argument as before, this results in the desired sub-cluster posterior distributions expressed in Equations 15-17.

C Hastings Ratios for Splits

In this section, we develop the Hastings ratio for a split proposal. We first note the following useful distribution

$$p(z)p(\pi|z) = \frac{\alpha^K \Gamma(\alpha) \prod_k \Gamma(N_k)}{\Gamma(\alpha + N)} \frac{\Gamma(\alpha + N)}{\Gamma(\alpha) \prod_k \Gamma(N_k)} \pi_{K+1}^\alpha \prod_k \pi_k^{N_k - 1} = \alpha^K \pi_{K+1}^\alpha \prod_{k=1}^K \pi_k^{N_k - 1}. \tag{18}$$

The Hastings ratio for a split can be expressed as

$$H_{\text{split-}c} = \frac{p(\hat{\pi}, \hat{z}, \hat{\theta}, x) p(\hat{\overline{\pi}}, \hat{\overline{\theta}}, \hat{\overline{z}}|x, \hat{z})}{p(\pi, z, \theta, x) p(\overline{\pi}, \overline{\theta}, \overline{z}|x, z)} \cdot \frac{q(\pi, z, \theta, \overline{\pi}, \overline{\theta}, \overline{z}|\hat{\pi}, \hat{z}, \hat{\theta}, \hat{\overline{\pi}}, \hat{\overline{\theta}}, \hat{\overline{z}}, Q_{\text{merge-}mn})}{q(\hat{\pi}, \hat{z}, \hat{\theta}, \hat{\overline{\pi}}, \hat{\overline{\theta}}, \hat{\overline{z}}|\pi, z, \theta, \overline{\pi}, \overline{\theta}, \overline{z}, Q_{\text{split-}c})}. \tag{19}$$

Because of the deferred proposal for auxiliary variables, this can be simplified to

$$H_{\text{split-}c} = \frac{p(\hat{\pi}, \hat{z}, \hat{\theta}, x)}{p(\pi, z, \theta, x)} \cdot \frac{q(\pi, z, \theta|\hat{\pi}, \hat{z}, \hat{\theta}, \hat{\overline{\pi}}, \hat{\overline{\theta}}, \hat{\overline{z}}, Q_{\text{merge-}mn})}{q(\hat{\pi}, \hat{z}, \hat{\theta}|\pi, z, \theta, \overline{\pi}, \overline{\theta}, \overline{z}, Q_{\text{split-}c})}. \tag{20}$$

We now analyze these terms separately.

The posterior ratio of z and π can be easily simplified to

$$\frac{p(\hat{z})p(\hat{\pi}|\hat{z})}{p(z)p(\pi|z)} = \frac{\alpha \hat{\pi}_{\hat{m}}^{\hat{N}_{\hat{m}} - 1} \hat{\pi}_{\hat{n}}^{\hat{N}_{\hat{n}} - 1}}{\pi_c^{N_c - 1}}. \tag{21}$$

This results in the following posterior ratio

$$\frac{p(\hat{\pi},\hat{z},\hat{\theta},x)p(\hat{\overline{\pi}},\hat{\overline{\theta}},\hat{\overline{z}}|x,\hat{z})}{p(\pi,z,\theta,x)p(\overline{\pi},\overline{\theta},\overline{z}|x,z)} = \boxed{\frac{\alpha\hat{\pi}_{\hat{m}}^{\hat{N}_{\hat{m}}-1}\hat{\pi}_{\hat{n}}^{\hat{N}_{\hat{n}}-1}}{\pi_c^{N_c-1}} \cdot \frac{f_\theta(\hat{\theta}_m;\lambda)f_x(x_{\{m\}};\hat{\theta}_m)\, f_\theta(\hat{\theta}_n;\lambda)f_x(x_{\{n\}};\hat{\theta}_n)}{f_\theta(\theta_c;\lambda)f_x(x_{\{k\}};\theta_k)}}. \tag{22}$$

The ratio of proposal distributions can similarly be simplified. We first note that the proposal for the new labels is an indicator function at the particular split or merge move:

$$q(\hat{z}_{\{m\}},\hat{z}_{\{n\}}|z,\overline{z},Q_{\text{split-}c}) = \mathbb{1}[(\hat{z}_{\{m\}},\hat{z}_{\{n\}}) = \text{split-}c(z,\overline{z})], \tag{23}$$

$$q(z_{\{c\}}|\hat{z},\hat{\overline{z}},Q_{\text{merge-}mn}) = \mathbb{1}[(z_{\{c\}} = \text{merge-}mn(\hat{z})]. \tag{24}$$

We note one important observation; the reverse label move that merges the two proposed clusters does not depend on auxiliary variables. This results in all split moves being exactly reversible by a merge move. The proposal ratio for proposed labels can then simplify to

$$\boxed{\frac{q(z_{\{c\}}|\hat{z},\hat{\overline{z}},Q_{\text{merge-}mn})}{q(\hat{z}_{\{m\}},\hat{z}_{\{n\}}|z,\overline{z},Q_{\text{split-}c})} = 1} \tag{25}$$

Conditioned on these new labels, we use the Reversible-Jump MCMC (RJMCMC) [2] algorithm to calculate the term for the π's. RJMCMC is a generalization of Metropolis-Hastings where auxiliary variables may be used to propose deterministic moves in mismatched dimensions. In general, if $x \in \mathbb{R}^D$ is the current D-dimensional variable and we desire to transition to $\hat{x} \in \mathbb{R}^{D+d}$, we augment the current space to be $[x, v]$ and the proposed space to be $[\hat{x}, \hat{u}]$. The auxiliary variables v and $\hat{u}$ must be chosen so that both spaces have the same dimensionality (i.e. $v \in \mathbb{R}^{C+d}$ and $u \in \mathbb{R}^C$, where C is any integer that satisfies $C + d \geq 0$ and $C \geq 0$). An RJMCMC algorithm then performs the following steps

1. Generate auxiliary variables for the current state:
$$v \sim q(v|x). \tag{26}$$
2. Apply a deterministic function, f to the current state to obtain a new proposal:
$$[\hat{x}, \hat{u}] = f(x, v). \tag{27}$$
3. Accept the proposal with the following probability
$$\min\left[1, \frac{p(\hat{x})q(\hat{u}|\hat{x})}{p(x)q(v|x)}\left|\det(J_f)\right|\right]. \tag{28}$$

Here, J_f is the Jacobian matrix of function $f(x, v)$. We will refer to the ratio, $\frac{q(\hat{u}|\hat{x})}{q(v|x)}\left|\det(J_f)\right|$ as the RJ proposal ratio, which generalizes the typical proposal ratio in the Metropolis-Hastings algorithm.

We now calculate the RJ proposal ratio of the π's using the RJCMCMC algorithm. The split proposal can be expressed as

$$[\pi_c, v] \rightarrow [\hat{\pi}_m, \hat{\pi}_n], \tag{29}$$

where $v \sim \text{Beta}(\hat{N}_m, \hat{N}_n)$ and the deterministic mapping between dimensions is

$$\hat{\pi}_m = \pi_c v, \quad \hat{\pi}_n = \pi_c(1 - v). \tag{30}$$

Because of the relationship between the Beta distribution and the Dirichlet distribution, the above construction can also be seen as a draw from a Dirichlet distribution followed by scaling by π_k. The Jacobian matrix can be expressed as

$$J_\pi = \begin{bmatrix} \frac{\partial\hat{\pi}_m}{\partial\pi_k} & \frac{\partial\hat{\pi}_m}{\partial v} \\ \frac{\partial\hat{\pi}_n}{\partial\pi_k} & \frac{\partial\hat{\pi}_n}{\partial v} \end{bmatrix} = \begin{bmatrix} v & \pi_c \\ (1-v) & -\pi_c \end{bmatrix}, \tag{31}$$

which has a corresponding absolute value determinant

$$|\det(J_\pi)| = |-\pi_c v - (\pi_c(1 - v))| = \pi_c. \tag{32}$$

Therefore, the RJ proposal ratio for the $\pi's$ can be expressed as

$$\frac{q(\pi|\hat{\bullet})}{q(\hat{\pi}|\bullet)} = \frac{1}{\text{Beta}(v;\hat{N}_m,\hat{N}_n)}\pi_c = \frac{\Gamma(\hat{N}_m)\Gamma(\hat{N}_n)}{\Gamma(N_c)} v^{\hat{N}_m-1}(1-v)^{\hat{N}_n-1}\pi_c \tag{33}$$

$$= \frac{\Gamma(\hat{N}_m)\Gamma(\hat{N}_n)}{\Gamma(N_c)} \left(\frac{\hat{\pi}_m}{\pi_c}\right)^{\hat{N}_m-1} \left(\frac{\hat{\pi}_n}{\pi_c}\right)^{\hat{N}_n-1} \pi_c \tag{34}$$

$$= \boxed{\frac{\Gamma(\hat{N}_m)\Gamma(\hat{N}_n)}{\Gamma(N_c)} \hat{\pi}_m^{\hat{N}_m-1}\hat{\pi}_n^{\hat{N}_n-1}\pi_c^{-N_c+1}}. \tag{35}$$

A similar reversible-jump proposal ratio can be found for the θ's. This requires one to define a 6-dimensional augmented space which we do not express here. Because the Jacobian can be written as the identity matrix, the determinant is simply 1. Thus, the ratio can be expressed as

$$\boxed{\frac{q(\theta|\hat{\bullet})}{q(\hat{\theta}|\bullet)} = \frac{q(\theta_c|x,z,\overline{z})}{q(\hat{\theta}_m|x,\hat{z},\hat{\overline{z}})q(\hat{\theta}_m|x,\hat{z},\hat{\overline{z}})}}. \tag{36}$$

Combining the boxed expressions, we arrive at the split Hastings ratio of Equation 23.

D Hastings Ratios for Merges

As stated in the paper, the Hastings ratio for a merge move is more complicated. This is due to the following proposal ratio for labels

$$\frac{q(z_{\{m\}}, z_{\{n\}}|\hat{z},\hat{\overline{z}},Q_{\text{split-}c})}{q(\hat{z}_{\{c\}}|z,\overline{z},Q_{\text{merge-}mn})}. \tag{37}$$

The numerator calculates the probability of splitting the proposed merged cluster back into the original two clusters. This means that the sub-cluster labels, $\hat{\overline{z}}$, must exactly correspond to the current clusters for the ratio to be non-zero. In practice, this is very unlikely for most situations. Consider the case where cluster m and cluster n have essentially the same weights and parameters (e.g. both Gaussian with the same mean and covariance). In the limit, this results in $p(z_{\{m\}}, z_{\{n\}}) = \left(\frac{1}{2}\right)^{N_m+N_n}$ since all configurations are the same. This means that with probability $\left(\frac{1}{2}\right)^{N_m+N_n}$, the proposed sub-cluster labels exactly correspond to the original regular-clusters. Clearly, this probability is very small for large N. When the proposed sub-cluster label does not correspond to the regular-clusters, the proposal is automatically rejected since $q(z_{\{m\}}, z_{\{n\}}|\hat{z},\hat{\overline{z}},Q_{\text{split-}c}) = 0$.

As the clusters m and n become more separable, the probability of proposing merged sub-cluster labels that exactly correspond to the regular-cluster labels increases. Imagine the limiting case where clusters m and n are infinitely far apart so that the only non-zero label assignment is the one that splits the data points correctly. In this case, the probability of the labels is one, and the probability of proposing the corresponding merged sub-cluster labels also is one. This results in a proposal ratio for labels equalling one. However, under the same assumption that the clusters m and n are very separable, it should be intuitive that they should *not* be merged to begin with. As such, the posterior ratio in the Hastings ratio will begin to dominate the overall acceptance ratio, and consequently still approach zero.

More precisely, the probability of accepting a proposed merge can be expressed as

$$\min[1, H_{\text{merge-}mn}]$$

$$= \min\left[1, \frac{\Gamma(\hat{N}_c)f_x(x_{\{c\}};\lambda)}{\alpha\prod_{k\in\{m,n\}}\Gamma(N_k)f_x(x_{\{k\}};\lambda)} \mathbb{1}\left[(\hat{\overline{z}}_{\{c,\ell\}}, \hat{\overline{z}}_{\{c,r\}}) = (z_{\{m\}}, z_{\{n\}})\right]\right] \tag{38}$$

$$= \min\left[1, \frac{\Gamma(\hat{N}_c)f_x(x_{\{c\}};\lambda)}{\alpha\prod_{k\in\{m,n\}}\Gamma(N_k)f_x(x_{\{k\}};\lambda)}\right] \mathbb{1}\left[(\hat{\overline{z}}_{\{c,\ell\}}, \hat{\overline{z}}_{\{c,r\}}) = (z_{\{m\}}, z_{\{n\}})\right]. \tag{39}$$

Again, we have assumed conjugate priors for simplifying the explanation.

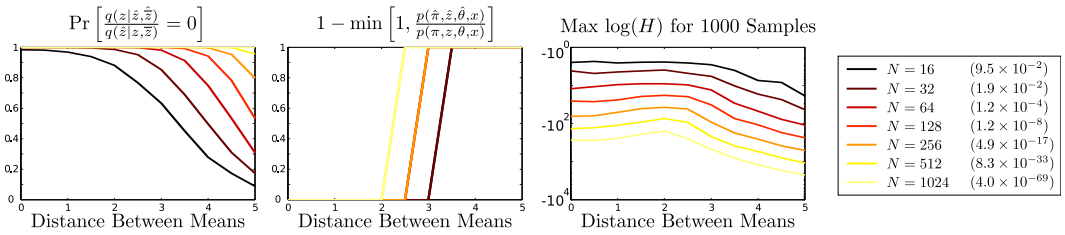

Figure 1: Probability quantities associated with rejecting a merge proposal. The numbers in the parenthesis correspond to the maximum observed upper bounds to the acceptance ratio over all samples and separations.

We note the following inequality

$$f_x(x_{\{c\}};\lambda) = \int f_x(x_{\{c\}};\theta_c) f_\theta(\theta_c;\lambda) d\theta_c \leq \int f_x(x_{\{c\}};\theta_c^*) f_\theta(\theta_c;\lambda) d\theta_c = f_x(x_{\{c\}};\theta_c^*), \quad (40)$$

where θ_c^* is the mode of the distribution and can be expressed as, $\theta_c^* = \max_{\theta_c} f_x(x_{\{c\}};\theta_c)$. For any realization of data, the probability of accepting a proposed merge can then be upper bounded by the following

$$\begin{aligned}
&\min[1, H_{\text{merge-}mn}] \\
&= \min\left[1, \frac{\Gamma(\hat{N}_c) f_x(x_{\{c\}};\lambda)}{\alpha \prod_{k\in\{m,n\}} \Gamma(N_k) f_x(x_{\{k\}};\lambda)}\right] \Pr\left[\hat{\tilde{z}}_{\{c\}} = (z_{\{m\}}, z_{\{n\}}) | x, z\right] \quad (41) \\
&\leq \min\left[1, \frac{\Gamma(\hat{N}_c) f_x(x_{\{c\}};\lambda)}{\alpha \prod_{k\in\{m,n\}} \Gamma(N_k) f_x(x_{\{k\}};\lambda)}\right] \Pr\left[\hat{\tilde{z}}_{\{c\}} = (z_{\{m\}}, z_{\{n\}}) | x, z, \overline{\theta}_c^*\right]. \quad (42)
\end{aligned}$$

We test our approximation that this acceptance ratio is 0 on synthetic data. We generate data from two 1D Gaussian distributions with mean separated by some amount. We use a Gaussian prior on the mean and assume the variance is known. The probability that the proposed merge is automatically rejected due to the second term in Equation 42 for varying separations and varying number of data points is shown in the first panel of Figure 1. Each value on each curve was calculated with 1000 samples of N data points. Clearly, as N increases, the proposed merge is rejected automatically more and more frequently. As expected, as the separation between the two clusters is increased, the automatic rejection is less likely.

Next, we show the probability of rejecting the merge due to the first term in Equation 42. The curves are shown in the second panel of Figure 1. This plot shows that as the separation increases, the posterior ratio favors rejecting the sample more and more. The overall log acceptance ratio is shown in the last panel of Figure 1 (note the additional log scale). In the legend, the values in the parenthesis indicate the maximum observed upper bounds to the acceptance ratio over all samples and separations. Even for relatively small N, approximating the proposals for merge moves with automatic rejections is very good. We note that the samplers are typically run for less than 10^3 iterations even for large datasets of $N \approx 10^6$, reinforcing the validity of the approximation even more.

E Random Splits and Merges

Since the merges are rejected automatically, we have included another pair of split/merge moves that typically propose good merges. As with any Metropolis-Hastings sampler, proposal distributions that are close to the true conditional distribution are desired. As such, we generate a random split of cluster c into clusters m and n as follows:

$$q_R(\hat{z}|z, Q_{\text{rsplit-}c}) = \text{Dir-Mult}(\hat{z}; \alpha/2, \alpha/2, N_c), \quad (43)$$

where notation is slightly abused, since only the data points assigned to cluster c are reassigned. Here, $Q_{\text{rsplit-}c}$ denotes the prior probability of proposing to split cluster c. To sample from the Dirichlet-Multinomials, we first sample temporary proportions, $\tilde{\pi}_m$ and $\tilde{\pi}_n$, from a Dirichlet distribution:

$$(\tilde{\pi}_m, \tilde{\pi}_n) \sim \text{Dir}(\alpha, \alpha), \quad (44)$$

followed by sampling the new assignments from a categorical distribution

$$\hat{z}_i \sim \text{Cat}(\tilde{\pi}_m, \tilde{\pi}_n), \quad \forall i \in \{i; z_i = c\}. \tag{45}$$

This is easily verified to be a valid sample from the proposal in Equation 43.

The corresponding random merge move is deterministic, since there is only one way to merge two clusters.

The Hastings ratio for a random split can then be expressed as

$$H_{\text{rsplit-}c} = \frac{p(\hat{z})p(x|\hat{z})}{p(z)p(x|z)} \frac{Q_{\text{rmerge-}mn}}{Q_{\text{rsplit-}c} \cdot \text{Dir-Mult}(\hat{z}; \alpha/2, \alpha/2, N_c)} \tag{46}$$

$$= \frac{\alpha\Gamma(\hat{N}_m)\Gamma(\hat{N}_n)}{\Gamma(N_c)} \frac{p(x|\hat{z})}{p(x|z)} \frac{Q_{\text{rmerge-}mn}}{Q_{\text{rsplit-}c}} \frac{\Gamma(\alpha + N_c)}{\Gamma(\alpha)} \frac{\Gamma(\alpha/2)^2}{\Gamma(\hat{N}_m)\Gamma(\hat{N}_n)} \tag{47}$$

$$= \frac{\alpha\Gamma(\alpha/2)^2\Gamma(\alpha + N_c)}{\Gamma(\alpha)\Gamma(N_c)} \frac{p(x|\hat{z})}{p(x|z)} \frac{Q_{\text{rmerge-}mn}}{Q_{\text{rsplit-}c}} \tag{48}$$

Similarly, the Hastings ratio for a random merge can then be expressed as

$$H_{\text{rmerge-}mn} = \frac{\Gamma(\alpha)\Gamma(N_m + N_n)}{\alpha\Gamma(\alpha/2)^2\Gamma(\alpha + N_m + N_n)} \frac{p(x|\hat{z})}{p(x|z)} \frac{Q_{\text{rsplit-}c}}{Q_{\text{rmerge-}mn}}. \tag{49}$$

We note that the Hastings ratio for a random merge proposal can be calculated efficiently from the summary statistics since the two current clusters, m and n, are already instantiated. Thus, a merge can be proposed in constant time. The Hastings ratio for a random split proposal depends on the resulting split cluster assignments, $\hat{z}$. Consequently, a random split proposal requires linear time in the size of the cluster. We therefore choose $Q_{\text{rsplit-}c} = 0.01 \times Q_{\text{rmerge-}mn}$ so that the random split proposals do not take too much computation. We note that any value besides 0.01 could be used without much difference.

References

[1] T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.

[2] P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995.