
Parallel Markov Chain Monte Carlo for Nonparametric Mixture Models: Supplementary Material

Sinead A. Williamson
Avinava Dubey
Eric P. Xing

SINEAD@CS.CMU.EDU
AKDUBEY@CS.CMU.EDU
EPXING@CS.CMU.EDU

Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15201, USA

In this document, we provide more in-depth proofs of the theorems and derive the Metropolis Hastings acceptance probabilities presented in the main paper.

1. Theorem expanded proofs

Theorem 1 (Auxiliary variable representation for the DPMM). *We can re-write the generative process for a DPMM as*

$$D_j \sim \text{DP}\left(\frac{\alpha}{P}, H\right), \quad \phi \sim \text{Dirichlet}\left(\frac{\alpha}{P}, \dots, \frac{\alpha}{P}\right), \quad (1)$$

$$\pi_i \sim \phi, \quad \theta_i \sim D_{\pi_i}, \quad x_i \sim f(\theta_i),$$

for $j = 1, \dots, P$ and $i = 1, \dots, N$. The marginal distribution over the x_i remains the same.

Proof. In the main paper, we proved the general result, that if $\phi \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_P)$ and $D_j \sim \text{DP}(\alpha_j, H_j)$, then $D := \sum_j \phi_j D_j \sim \text{DP}(\sum_j \alpha_j, \frac{\sum_j \alpha_j H_j}{\sum_j \alpha_j})$. This result has been used by authors including Rao & Teh (2009).

Here, we provide an explicit proof that shows the resulting predictive distribution is that of the Dirichlet process.

Let $\theta_1, \theta_2, \dots$ be a sequence of random variable distributed according to $G \sim \text{DP}(\alpha, G_0)$. Then the conditional distribution of θ_{n+1} given $\theta_1, \dots, \theta_n$ where G has been integrated is given by

$$\theta_{n+1} | \theta_1, \dots, \theta_n \sim \sum_{l=1}^n \frac{1}{n+\alpha} \delta_{\theta_l} + \frac{\alpha}{n+\alpha} G_0. \quad (2)$$

If $D_j \sim \text{DP}(\alpha/P, G_0)$, $\phi \sim \text{Dir}(\frac{\alpha}{P}, \dots, \frac{\alpha}{P})$, $\pi_i \sim \phi$ and $\theta_i \sim D_{\pi_i}$ then the conditional distribution of θ_{n+1} given $\theta_1, \dots, \theta_n$ where $D_j, \forall j$ and ϕ have been inte-

grated is given by

$$\begin{aligned} & \theta_{n+1} | \theta_1, \dots, \theta_n \\ & \sim \sum_{j=1}^P P(\pi_{n+1} = j | \pi_1, \dots, \pi_n) \\ & \quad \cdot P(\theta_{n+1} | \pi_{n+1} = j, \pi_1, \dots, \pi_n, \theta_1, \dots, \theta_n, G_0) \\ & = \sum_j \frac{n_j + \alpha/P}{n + \alpha} \\ & \quad \cdot \left\{ \sum_{l=1}^n \frac{1}{n_j + \alpha/P} \delta_{\theta_l} \delta_{\pi_l=j} + \frac{\alpha/P}{n_j + \alpha/P} G_0 \right\} \\ & = \sum_{l=1}^n \frac{1}{n + \alpha} \delta_{\theta_l} + \frac{\alpha}{n + \alpha} G_0. \end{aligned} \quad (3)$$

□

Theorem 2 (Auxiliary variable representation for the HDP). *If we incorporate the requirement that the concentration parameter γ for the bottom level DPs $\{D_j\}_{j=1}^M$ depends on the concentration parameter α for the top level DP D_0 as $\gamma \sim \text{Gamma}(\alpha)$, then we can rewrite the generative process for the HDP as:*

$$\begin{aligned} \zeta_j & \sim \text{Gamma}(\alpha/P), & \pi_{mi} & \sim \nu_m, \\ D_{0j} & \sim \text{DP}(\alpha/P, H), & \theta_{mi} & \sim D_{m\pi_{mi}}, \\ \nu_m & \sim \text{Dirichlet}(\zeta_1, \dots, \zeta_P), & x_{mi} & \sim f(\theta_{mi}), \\ D_{mj} & \sim \text{DP}(\zeta_j, D_{0j}), \end{aligned} \quad (4)$$

for $j = 1, \dots, P$, $m = 1, \dots, M$, and $i = 1, \dots, N_m$.

Proof. Let $\zeta_j \sim \text{Gamma}(\alpha/P)$ and $D_{0j} \sim \text{DP}(\alpha/P, H)$, $j = 1, \dots, P$. This implies that $G_{0j} := \zeta_j D_{0j} \sim \text{GaP}((\alpha/P)H)$ and $\gamma := \sum_{j=1}^P \zeta_j \sim \text{Gamma}(\alpha)$.

By superposition of gamma processes,

$$\begin{aligned} G_0 &:= \sum_{j=1}^P G_{0j} \sim \text{GaP}(\alpha H) \\ &:= \sum_{j=1}^P \zeta_j D_{0j}. \end{aligned}$$

Normalizing G_0 , we get

$$D_0(\cdot) := \frac{G_0(\cdot)}{\gamma} = \sum_{j=1}^P \frac{\zeta_j}{\gamma} D_{0j} \sim \text{DP}(\alpha, H)$$

as required by the HDP.

Now, for $m = 1, \dots, M$ and $j = 1, \dots, P$, let $\eta_{mj} \sim \Gamma(\zeta_j)$ and $D_{mj} \sim \text{DP}(\zeta_j, D_{0j})$. This implies that

$$\begin{aligned} G_{mj} &:= \eta_{mj} D_{mj} \sim \text{GaP}(\zeta_j D_{0j}) \\ &:= \text{GaP}(G_{0j}). \end{aligned}$$

Superposition of the gamma processes gives

$$\begin{aligned} G_m &:= \sum_j G_{mj} \sim \text{GaP}\left(\sum_{j=1}^P G_{0j}\right) \\ &:= \text{GaP}(G_0) = \text{GaP}(\gamma D_0). \end{aligned}$$

The total mass of G_m is given by $\sum_{j=1}^P \eta_{mj}$, so

$$D_m(\cdot) = \frac{G_m(\cdot)}{\sum_{j=1}^P \eta_{mj}} \sim \text{DP}(\gamma, D_0) \quad (5)$$

as required by the HDP.

If we let $\nu_{mj} = \eta_{mj} / \sum_{k=1}^P \eta_{mk}$, then we can rewrite Equation 5 as

$$D_m(\cdot) = \sum_{j=1}^P \nu_{mj} D_{mj} \sim \text{DP}(\gamma, D_0), \quad (6)$$

where $(\nu_{m1}, \dots, \nu_{mP}) \sim \text{Dirichlet}(\zeta_1, \dots, \zeta_P)$. \square

2. Metropolis Hastings acceptance probabilities

In both algorithms, the Metropolis Hastings proposal probabilities satisfy $q(\{\pi_i\} \rightarrow \{\pi_i^*\}) = q(\{\pi_i^*\} \rightarrow \{\pi_i\})$, so we need only consider the likelihood ratios.

2.1. Dirichlet process

In the Dirichlet process case, the likelihood ratio is given by:

$$\begin{aligned} \frac{p(\{\pi_i^*\})}{p(\{\pi_i\})} &= \frac{p(\{x_i\}|\pi_i^*)p(\{\pi_i^*\}|\alpha, P)}{p(\{x_i\}|\pi_i)p(\{\pi_i\}|\alpha, P)} \\ &= \frac{p(\{z_i\}|\pi_i^*)p(\{\pi_i^*\}|\alpha, P)}{p(\{z_i\}|\pi_i)p(\{\pi_i\}|\alpha, P)} \quad (7) \\ &= \prod_{j=1}^P \prod_{i=1}^{\max(N_j, N_j^*)} \frac{a_{ij}!}{a_{ij}^*!}, \end{aligned}$$

where N_j is the number of data points on processor j , and a_{ij} is the number of clusters of size i on processor j .

The probability of the processor allocations is described by the Dirichlet compound multinomial, or multivariate Pólya, distribution,

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

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

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

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 k th on processor j . The distribution over cluster sizes in the Chinese restaurant process is described by Ewen's sampling formula, which gives:

$$p(\{n_{jk}\}|N_j) = \binom{\alpha}{P}^{K_j} \frac{N_j!}{\prod_{k=1}^{K_j} n_{jk}!} \frac{\Gamma(\alpha/P)}{\Gamma(N_j + \alpha/P)} \prod_{i=1}^{N_j} \frac{1}{a_j!}$$

where K_j is the total number of clusters on processor j . Therefore,

$$\frac{p(\{n_{jk}^*\}|N_j)}{p(\{n_{jk}\}|N_j)} = \prod_{j=1}^P \frac{N_j^*!}{N_j!} \frac{\Gamma(N_j + \alpha/P)}{\Gamma(N_j^* + \alpha/P)} \prod_{i=1}^{\max(N_j, N_j^*)} \frac{a_{ij}!}{a_{ij}^*!},$$

so we get Equation 7.

2.2. Hierarchical Dirichlet processes

For the HDP, the likelihood ratio is given by

$$\frac{p(\{x_{mi}\}|\{\pi_{mi}^* \gamma, \boldsymbol{\xi}^*, \alpha, P\}) p(\{\pi_{mi}^*\}|\gamma, \boldsymbol{\xi}^*) p(\boldsymbol{\xi}^*|\alpha, P)}{p(\{x_{mi}\}|\{\pi_{mi} \gamma, \boldsymbol{\xi}, \alpha, P\}) p(\{\pi_{mi}\}|\gamma, \boldsymbol{\xi}) p(\boldsymbol{\xi}|\alpha, P)}. \quad (8)$$

We consider a Chinese restaurant franchise representation (Teh et al., 2006), 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 DP).

Let \mathbf{t}_j be the count vector for the top-level DP 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 m th bottom-level DP on processor j – in Chinese restaurant franchise terms, n_{jmk} is the number of customers in the m th restaurant sat at the k th table of the j th processor. Let T_{mj} be the total number of occupied tables from the m th 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. 8 can be re-written as:

$$\frac{p(\{t_{jd}\}, \{n_{jmk}^*\}|\{\pi_{mi}^* \gamma, \boldsymbol{\xi}^*, \alpha, P\})}{p(\{t_{jd}\}, \{n_{jmk}\}|\{\pi_{mi} \gamma, \boldsymbol{\xi}, \alpha, P\})} \cdot \frac{p(\{\pi_{mi}^*\}|\gamma, \boldsymbol{\xi}^*) p(\boldsymbol{\xi}^*|\alpha, P)}{p(\{\pi_{mi}\}|\gamma, \boldsymbol{\xi}) p(\boldsymbol{\xi}|\alpha, P)}. \quad (9)$$

The first term in the Eq. 9 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- 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

$$p(\{t_{jd}\}|\alpha, P) = \prod_{j=1}^P \left(\frac{\alpha}{P}\right)^{U_j} \frac{T_{\cdot j}!}{\prod_{d=1}^{U_j} t_{jd}!} \frac{\Gamma(\alpha/P)}{\Gamma(\alpha/P + T_{\cdot j})} \prod_{i=1}^{T_{\cdot j}} \frac{1}{b_{ji}},$$

so

$$\begin{aligned} & \frac{p(\{t_{jd}^*\}, \{n_{jmk}^*\}|\{\pi_{mi}^* \gamma, \boldsymbol{\xi}^*, \alpha, P\})}{p(\{t_{jd}\}, \{n_{jmk}\}|\{\pi_{mi} \gamma, \boldsymbol{\xi}, \alpha, P\})} \\ &= \prod_{j=1}^P \frac{(\xi_j^*)^{T_{\cdot j}^*} T_{\cdot j}^*! \Gamma(\alpha/P + T_{\cdot j})}{(\xi_j)^{T_{\cdot j}} T_{\cdot j}! \Gamma(\alpha/P + T_{\cdot j}^*)} \left(\frac{\Gamma(\gamma \xi_j^*)}{\Gamma(\gamma \xi_j)}\right)^M \\ & \cdot \left\{ \prod_{i=1}^{\max(T_{\cdot j}, T_{\cdot j}^*)} \frac{b_{ji}!}{b_{ji}^*!} \right\} \prod_{m=1}^M \frac{n_{jm\cdot}^*! \Gamma(\gamma \xi_j + n_{jm\cdot})}{n_{jm\cdot}! \Gamma(\gamma \xi_j^* + n_{jm\cdot}^*)} \\ & \cdot \prod_{i=1}^{\max(N_j, N_j^*)} \frac{a_{jmi}!}{a_{jmi}^*!}. \end{aligned} \quad (10)$$

The probability of the processor assignments is given by:

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

so the second term is given by

$$\frac{p(\{\pi_{mi}^*\}|\gamma, \boldsymbol{\xi}^*)}{p(\{\pi_{mi}\}|\gamma, \boldsymbol{\xi})} = \prod_{j=1}^P \left(\frac{\Gamma(\gamma \xi_j)}{\Gamma(\gamma \xi_j^*)}\right)^M \prod_{m=1}^M \frac{n_{jm\cdot}! \Gamma(\gamma \xi_j^* + n_{jm\cdot}^*)}{n_{jm\cdot}^*! \Gamma(\gamma \xi_j + n_{jm\cdot})}. \quad (11)$$

The third term is given by

$$\frac{p(\boldsymbol{\xi}^*|\alpha, P)}{p(\boldsymbol{\xi}|\alpha, P)} = \prod_{j=1}^P \left(\frac{\xi_j^*}{\xi_j}\right)^{\frac{\alpha}{P}}. \quad (12)$$

Combining Equations 10, 11 and 12 gives an acceptance probability of $\min(1, r)$, where

$$r = \prod_{j=1}^P \frac{(\xi_j^*)^{T_{\cdot j}^* + \alpha/P} T_{\cdot j}^*! \Gamma(\alpha/P + T_{\cdot j})}{(\xi_j)^{T_{\cdot j} + \alpha/P} T_{\cdot j}! \Gamma(\alpha/P + T_{\cdot j}^*)} \prod_{i=1}^{n_{\cdot j}} \frac{b_{ji}!}{b_{ji}^*!} \prod_{m=1}^M \frac{a_{jmi}!}{a_{jmi}^*!}. \quad (13)$$

2.3. Sampling γ

We sample the HDP parameter γ using reversible random walk Metropolis Hastings steps, giving an acceptance probability of

$$\min\left(1, \left(\frac{\gamma^*}{\gamma}\right)^{T_{\cdot}} \left[\frac{\Gamma(\gamma^*)}{\Gamma(\gamma)}\right]^M \prod_{m=1}^M \frac{\Gamma(n_{m\cdot} + \gamma)}{\Gamma(n_{m\cdot} + \gamma^*)}\right).$$

References

- Rao, V. and Teh, Y.-W. Spatial normalized gamma processes. In *NIPS*, 2009.
- Teh, Y.-W., Jordan, M. I., Beal, M. J., and Blei, D. M. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.